ELECTRICAL ENGINEERS' HANDBOOK

Electric Communication and Electronics

WILEY ENGINEERING HANDBOOK SERIES

HANDBOOK OF ENGINEERING FUNDAMENTALS. Edited by OVID W. ESHBACH.

KENT'S MECHANICAL ENGINEERS' HANDBOOK. Twelfth Edition.

DESIGN AND PRODUCTION. Edited by COLIN CARMICHAEL.

POWER. Edited by J. K. SALISBURY.

ELECTRICAL ENGINEERS' HANDBOOK. Fourth Edition.

ELECTRIC POWER. Edited by HAROLD PENDER and WILLIAM A. DEL MAR.

ELECTRIC COMMUNICATION AND ELECTRONICS. Edited by HAROLD PENDER and KNOX MCILWAIN.

MINING ENGINEERS' HANDBOOK. Third Edition. Edited by the late ROBERT PEELE.

HANDBOOK OF MINERAL DRESSING.

ORES AND INDUSTRIAL MINERALS. Edited by ARTHUR F. TAGGART.

ELECTRICAL ENGINEERS' HANDBOOK

Electric Communication and Electronics

Prepared by a Staff of Specialists

HAROLD PENDER, PH.D., SC.D.

and

KNOX McILWAIN, B.S., E.E.

Editors

FOURTH EDITION

WILEY ENGINEERING
HANDBOOK SERIES

NEW YORK : JOHN WILEY & SONS, INC.

LONDON : CHAPMAN & HALL, LIMITED

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

The first edition of Pender's *Handbook for Electrical Engineers*, compiled by a staff of specialists under the editorship of Harold Pender, appeared in 1914. The second edition, under the joint editorship of Pender and William A. Del Mar, was published in 1922. Both these editions covered all branches of electrical engineering as well as a large amount of material dealing with allied fields of interest to electrical engineers.

The third edition, published in 1936, was divided into two volumes: one on electric power under the editorship of Pender, Del Mar, and Knox McIlwain; the other on electrical communication and electronics under the editorship of Pender and McIlwain. Certain tables and fundamental theory were duplicated in the two volumes in order that each might be complete and independent of the other.

This plan met with such enthusiastic response that it has been continued in the fourth edition. The growth of knowledge and the greater degree of specialization in the various phases of electrical engineering have necessitated a considerable enlargement of both volumes. Careful selection and compression have been exercised in an effort to keep the books compact and readable. The treatment of subjects of decreased importance and those which are adequately treated by other handbooks of the Wiley Handbook Series has been either curtailed or left unchanged in length.

The bibliographies have been prepared with the idea of assisting the reader to further study of each subject; and they reflect each author's idea of this plan. The publications referred to are, in general, in the Engineering Societies Library, 29 West 39th Street, New York, N. Y. Most of them may be borrowed from the Library by members of its Founder Societies, the American Society of Civil Engineers, American Institute of Mining and Metallurgical Engineers, American Society of Mechanical Engineers, and American Institute of Electrical Engineers.

Seventy-eight specialists in their respective fields have contributed to this fourth and entirely rewritten edition of the Electronics and Communication portion of the *Electrical Engineers' Handbook*, as compared with twenty-seven, forty-five, and fifty-seven in previous editions. This reflects the rapid widening in the electronics field. In particular, frequency modulation and all the pulse techniques in both the communication and radar fields appear in the volume for the first time. The increased complexity and importance of radio aids to navigation are also of interest.

The editors' thanks are due to the many well-known and busy men who have contributed textual material, both for their unselfish efforts to make this a reliable reference work and for their continued patience with editorial vagaries. They are also due to Messrs. R. L. Jones, R. K. Honaman, and A. R. Thompson of the Bell Telephone Laboratories, Mr. Frank A. Cowan of the American Telephone and Telegraph Company, and Mr. E. W. Engstrom of RCA Laboratories for aid in the organization of the book.

HAROLD PENDER
KNOX MCILWAIN

LIST OF CONTRIBUTORS

Maurice Artzt, Research Engineer, Radio Corp. of America, RCA Laboratories Division. *Facsimile.*

W. F. Bailey, Sr. Television Engineer, Hazeltine Electronics Corp. *Television Receivers.*

Loy E. Barton, Research Engineer, Radio Corp. of America, RCA Laboratories Division. *Amplifiers.*

Dr. R. M. Bozorth, Research Physicist, Bell Telephone Laboratories. *Magnetic Materials.*

Dr. J. G. Brainerd, Professor, Moore School of Electrical Engineering, University of Pennsylvania. *Units and Conversion Factors, Symbols and Abbreviations.*

R. B. J. Brunn, Sr. Engineer, Hazeltine Electronics Corp. *Television Receivers.*

T. J. Buzalski, Television Station Engineer, National Broadcasting Company. *Television Broadcasting.*

Dr. Carl C. Chambers, Acting Dean, Moore School of Electrical Engineering, University of Pennsylvania. *Mathematics, Mathematical Tables, Units and Conversion Factors, Symbols and Abbreviations, Constants, Oscillators.*

R. A. Chegwidden, Member of Technical Staff, Bell Telephone Laboratories. *Magnetic Materials.*

H. A. Chinn, Chief Audio Engineer, Columbia Broadcasting System. *Audio Facilities for Sound Systems, Radio Telephone Broadcasting.*

E. L. Clark, Sr. Television Engineer, Radio Corp. of America, RCA Victor Division. *Special Purpose Amplifiers.*

James I. Cornell, Vickers Electric Division, Vickers, Inc. *Capacitors, Condensers.*

L. F. Curtis, Consultant, Hazeltine Electronics Corp. *F-m Receivers.*

B. J. Dalton, Industrial Engineering Division, Apparatus Dept., General Electric Company. *Electronic Control Equipment.*

Paul S. Darnell, Transmission Apparatus Engineer, Bell Telephone Laboratories. *Resistors and Rheostats.*

Howard L. Davis, Jr., Engineer in Charge, Special Investigation Section, Engineering Department, Philadelphia Electric Company. *Coordination of Communication and Power Systems.*

William R. Dohan, General Consultant, Consulting and Standards Division, Engineering Dept., Radio Corp. of America, RCA Victor Division. *Insulating Materials.*

R. D. Duncan, Jr., Engineering Consultant. *F-m Systems.*

Dr. D. W. Epstein, Research Engineer, Radio Corp. of America, RCA Laboratories Division. *Geometrical Optics, Electron Optics.*

J. G. Ferguson, Engineer, Bell Telephone Laboratories. *Measurement of Primary Electrical Quantities.*

J. L. Finch, Plant Design Superintendent, Radio Corp. of America, RCA Communications, Inc. *Radio Telegraph Systems.*

H. J. Fisher, Test Engineer, Bell Telephone Laboratories. *Wire Line Measurements.*

F. J. Gaffney, Chief Engineer, Polytechnic Research and Development Co., Inc. *Signal Generators and Power Measurements.*

Dr. Maxwell K. Goldstein, Air Navigation Development Board, CAA, Dept. of Commerce. *Radio Aids to Marine Navigation.*

A. J. Grossman, Member of Technical Staff, Bell Telephone Laboratories. *Electric Wave Filters.*

O. B. Gunby, Recording Equipment Systems Engineer, Radio Corp. of America, RCA Victor Division. *Sound Recording.*

A. L. Hammerschmidt, Development Engineer, National Broadcasting Company. *Television Broadcasting.*

W. B. Hebenstreit, Section Head, Navigation and Radar Research, Electronics Department, Hughes Aircraft Company; formerly with Bell Telephone Laboratories. *Magnetrons.*

H. A. Henning, Member of Technical Staff, Bell Telephone Laboratories. *Magnetic Recording and Reproducing of Sound, Mechanical Recording and Reproducing of Sound.*

L. M. Hershey, Director of Research, General Instrument Corporation. *Inductors with Air Cores.*

C. J. Hirsch, Chief Engineer, Research Division, Hazeltine Electronics Corp. *I-f Amplifiers.*

E. W. Houghton, Engineer, Bell Telephone Laboratories. *Microwave Measurements.*

Dr. Herbert E. Ives, Research Consultant, Bell Telephone Laboratories. *Photoresponsive Devices.*

A. P. Kauzmann, Engineer, Vacuum Tube Advanced Development, Radio Corp. of America, RCA Victor Division. *Thermionic Vacuum Tubes.*

C. R. Keith, Engineering Representative, ERP Division, Western Electric Company. *Photographic Sound Recording.*

J. P. Kinzer, Member of the Technical Staff, Bell Telephone Laboratories. *Cavity Resonators.*

Hugh S. Knowles, Vice President, Jensen Manufacturing Company, President Industrial Research Products, Inc. *Effects of the Acoustic Medium, Loudspeakers and Telephone Receivers, Microphones.*

Vern O. Knudson, Dean and Professor of Physics, Graduate Division, University of California. *Acoustic Properties of Rooms, Sound Insulation, Acoustic Design of Auditoriums.*

R. J. Kowalski, Commercial Engineer, Custom Built Dept., Radio Corp. of America, RCA Victor Division. *Electro-acoustic Equipment, Public-address Systems.*

V. D. Landon, Radio Research Engineer, Radio Corp. of America, RCA Laboratories Division. *Single Mesh and Coupled Circuits, Detectors, Radio Receivers.*

H. W. Leverenz, Research Chemico-Physicist, Radio Corp. of America, RCA Laboratories Division. *Luminescent and Tenebrescent Materials.*

B. D. Loughlin, Sr. Engineer, Hazeltine Electronics Corp. *Distortion and Interference in F-m Systems.*

A. V. Loughren, Vice President in Charge of Research, Hazeltine Electronics Corp. *Television Principles and Theory.*

Warren A. Marrison, Member of Technical Staff, Bell Telephone Laboratories. *Frequency Measurements.*

Dr. Warren P. Mason, Research Physicist, Bell Telephone Laboratories. *Piezoelectric Crystals.*

Knox McIlwain, Chief Consulting Engineer, Hazeltine Electronics Corp. *Non-linear Electric Circuits, Electromagnetic Radiation, Electromechanical Systems, Conducting Materials, Single Mesh and Coupled Circuits.*

Henry I. Metz, Chief, Electronics Maintenance Branch, Maintenance Engineering Div., Civil Aeronautics Administration. *Radio Aids to Air Navigation.*

H. F. Mickel, Communication Sales Manager, Raymond Rosen Engineering Products, Inc. *Police Radio.*

W. A. Munson, Telephone Engineer, Bell Telephone Laboratories. *The Sense of Hearing, Speech and Music, Effect of Distortion.*

Dr. Kenneth N. Ogle, Optical Research Dept., Mayo Clinic. *Vision.*

J. J. Okrent, Sr. Engineer, Hazeltine Electronics Corp. *Pulse Circuits.*

D. S. Peck, Member Technical Staff, Bell Telephone Laboratories. *Gaseous Conduction Tubes.*

Harold Pender, Consultant, Moore School of Electrical Engineering, University of Pennsylvania.

Dr. E. Peterson, Member Technical Staff, Bell Telephone Laboratories. *Wave Analysis.*

J. D. Phyfe, Product Engineer, Theatre Sound Equipment, Radio Corp. of America, RCA Victor Division. *Projection Practices.*

N. Y. Priessman, Member of Technical Staff, Bell Telephone Laboratories. *Varistors and Thermistors.*

Dr. G. L. Ragan, Research Associate, General Electric Company. *Wave-guide Components.*

Dr. E. G. Ramberg, Research Physicist, Radio Corp. of America, RCA Laboratories Division. *Television Pick-up Tubes.*

P. H. Richardson, Member of Technical Staff, Bell Telephone Laboratories. *Theory of Linear Passive Networks, Recurrent Networks.*

Arnold J. Rohner, Project Engineer, Bendix Radio Division. *Ferrous Cored Inductors, Transformers with Iron Cores.*

A. L. Samuel, Professor of Electrical Engineering, University of Illinois. *Klystrons.*

Arthur H. Schafer, Member of Technical Staff, Bell Telephone Laboratories. *Resistors and Rheostats.*

Dr. S. A. Schelkunoff, Member of Technical Staff, Bell Telephone Laboratories. *Wave-guide Theory.*

J. C. Schelleng, Radio Research Engineer, Bell Telephone Laboratories. *Radio Antennas, Transmission in Space.*

Dr. Theodore Soller, Professor of Physics, Amherst College. *Cathode-ray Tube Displays.*

F. J. Somers, Staff Engineer, National Broadcasting Company. *Television Broadcasting.*

Dr. John C. Steinberg, Member of Technical Staff, Bell Telephone Laboratories. *The Sense of Hearing, Speech and Music, Effects of Distortion.*

L. E. Swedlund, Sr. Engineer, Radio Corp. of America, RCA Victor Division. *Cathode-ray Tubes.*

W. O. Swinyard, Chief Engineer, Hazeltine Research, Inc. *Routine Measurement of A-m and F-m Broadcast Receivers.*

John D. Taylor, Consulting Engineer, American Telephone & Telegraph Company. *Telephony, Telegraphy, Coordination of Communication and Power Systems, Wire Transmission Lines, Mechanical Features of Transmission Lines.*

L. Vieth, Member of Technical Staff, Bell Telephone Laboratories. *Magnetic Recording and Reproducing of Sound, Mechanical Recording and Reproducing of Sound.*

Arthur H. Volz, Member of Technical Staff, Bell Telephone Laboratories. *Potentiometers and Rheostats.*

Dr. S. Reid Warren, Jr., Associate Professor, Moore School of Electrical Engineering, Associate Professor Radiologic Physics, Graduate School of Medicine, University of Pennsylvania. *X-ray Tubes, Medical Applications of Electricity.*

Charles Weyl, Professor, Moore School of Electrical Engineering, Associate Professor of Radiologic Physics, Graduate School of Medicine, University of Pennsylvania. *Medical Applications of Electricity.*

H. A. Wheeler, Consulting Radio Physicist, engaged for this work by Hazeltine Electronics Corp. *Pulse Techniques, Transient Networks.*

I. G. Wilson, Member of Technical Staff, Bell Telephone Laboratories. *Cavity Resonators.*

J. E. Young, Supervisor, Broadcast Engineering Section, Radio Corp. of America, RCA Victor Division. *Modulators, Power Supply, Radio Transmitters, F-m Transmitters.*

Dr. V. K. Zworykin, Vice President and Technical Consultant, Radio Corp. of America, RCA Laboratories Division. *Television Pick-up Tubes.*

GENERAL TABLE OF CONTENTS

Detailed tables of contents are given at the beginning of each section. An alphabetical index appears after Section 23.

SECTION 16. SOUND-REPRODUCTION SYSTEMS

SECTION 17. TELEPHONY

SECTION 18. TELEGRAPHY

SECTION 19. FACSIMILE TRANSMISSION AND RECEPTION

SECTION 20. TELEVISION

SECTION 21. ELECTRONIC CONTROL EQUIPMENT

SECTION 22. AIDS TO NAVIGATION

SECTION 23. MEDICAL APPLICATIONS OF ELECTRICITY

This book is divided into sections, each section carrying its independent sequence of page numbers. For example, 3-15 indicates Section 3, page 15.

SECTION 1

MATHEMATICS, UNITS, AND SYMBOLS

MATHEMATICS

By Carl C. Chambers

MATHEMATICAL TABLES AND CHARTS

UNITS AND CONVERSION FACTORS

By J. G. Brainerd and Carl C. Chambers

SYMBOLS AND ABBREVIATIONS

CONSTANTS

By Carl C. Chambers

MATHEMATICS, UNITS, AND SYMBOLS

MATHEMATICS

By Carl C. Chambers

1. ALGEBRAIC FORMULAS

MISCELLANEOUS FORMULAS

$$(a \pm b)^2 = a^2 \pm 2ab + b^2$$

$$(a \pm b)^3 = a^3 \pm 3a^2 b + 3ab^2 \pm b^3$$

$$(a \pm b)^n = \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} a^k (\pm b)^{(n-k)}, \qquad n! = n(n-1)\ldots 3 \times 2 \times 1$$

$$a^2 - b^2 = (a+b)(a-b)$$

$$a^2 + b^2 = (a+jb)(a-jb), \quad j = \sqrt{-1}$$

$$a^x \times a^y = a^{(x+y)}, \quad a^\circ = 1 \text{ [for } a \neq 0], \quad (ab)^x = a^x b^x$$

$$\frac{a^x}{a^y} = a^{(x-y)}, \quad a^{-x} = \frac{1}{a^x}, \quad \left(\frac{a}{b}\right)^x = \frac{a^x}{b^x}$$

$$(a^x)^y = a^{xy}, \quad a^{1/x} = \sqrt[x]{a}, \quad \sqrt[x]{ab} = \sqrt[x]{a}\,\sqrt[x]{b}$$

$$\sqrt[x]{\sqrt[y]{a}} = \sqrt[xy]{a}, \quad a^{x/y} = \sqrt[y]{a^x}, \quad \sqrt[x]{\frac{a}{b}} = \frac{\sqrt[x]{a}}{\sqrt[x]{b}}$$

$$\log (a^x) = x \log a, \quad \log ab = \log a + \log b$$

$$\log \frac{a}{b} = \log a - \log b$$

$$\text{If } \frac{a}{b} = \frac{c}{d} \text{ then } \frac{a \pm b}{b} = \frac{c \pm d}{d} \text{ and } \frac{a-b}{a+b} = \frac{c-d}{c+d}$$

The sum of an arithmetical progression is given by

$$s = \frac{n}{2}(a + l) = \frac{n}{2}\{2a + (n-1)d\}$$

where $l = a + (n-1)d$ is the last term, a is the first term, d is the common difference, and s is the sum of the n terms.

The sum of a geometrical progression is given by

$$s = a\,\frac{(1 - r^n)}{1 - r} = \frac{lr - a}{r - 1}$$

where $l = ar^{n-1}$ is the last term, a is the first term, d is the common ratio, and s is the sum of the n terms. If n approaches infinity and r^2 is less than unity

$$s = \frac{a}{1 - r}$$

The multiple product represented by $n(n-1)(n-2)\ldots 3 \times 2 \times 1$ is designated by the symbol $n!$ or $\underline{|n}$ and is called " n factorial." The following list gives the value of $n!$ up to $n = 10$

1! =	1	6! =	720
2! =	2	7! =	5,040
3! =	6	8! =	40,320
4! =	24	9! =	362,880
5! =	120	10! =	3,628,800

For large values of n a good approximation for $n!$ is, from Stirling's formula,

$$n! = (2\pi n)^{1/2} \left(\frac{n}{e}\right)^n, \quad e = 2.7182818$$

This formula is accurate to about $2^1/_2$ per cent at $n = 10$ and becomes more accurate very rapidly as n is increased,

The number of permutations or arrangements of n things taken p at a time is

$$P_p^n = \frac{n!}{(n-p)!}$$

The number of combinations of n things taken p at a time is then

$$C_p^n = \frac{1}{p!} P_p^n$$

QUADRATIC EQUATION. The solution of

$$ax^2 + bx + c = 0$$

is

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

If a, b, and c are real, and the discriminant, $b^2 - 4ac$, is positive, the roots are real and unequal; if it is zero, the roots are real and equal; if it is negative, the roots are conjugate complex numbers.

CUBIC EQUATIONS. The solution of

$$ax^3 + 3bx^2 + 3cx + d = 0 \qquad (1)$$

is obtained as follows: Put $x = \frac{1}{a}(y - b)$; then (1) becomes

$$y^3 - 3Hy + G = 0$$

where

$$H = b^2 - ac$$
$$G = a^2 d - 3abc + 2b^3$$

For a solution let

$$A = \sqrt[3]{\frac{-G}{2} + \sqrt{\frac{G^2}{4} - H^3}}, \quad B = \sqrt[3]{\frac{-G}{2} - \sqrt{\frac{G^2}{4} - H^3}}$$

then the values of y will be given by

$$y = A + B, \quad -\frac{1}{2}(A + B) + j\frac{\sqrt{3}}{2}(A - B), \quad -\frac{1}{2}(A + B) - j\frac{\sqrt{3}}{2}(A - B)$$

If a, b, c, d are real and if $G^2 - 4H^3$, the discriminant, is positive there are one real root and two conjugate complex roots; if $G^2 - 4H^3$ is zero there are three real roots, at least two of which are equal; if $G^2 - 4H^3$ is negative there are three real and unequal roots.

The solution may be written in three other forms.

(1) Put

$$\phi = \frac{1}{3}\sin^{-1}\left[\frac{G}{2H\sqrt{H}}\right]$$

then the roots are

$$y = 2\sqrt{H}\sin\phi, \quad 2\sqrt{H}\sin(\phi + 120°), \quad 2\sqrt{H}\sin(\phi - 120°)$$

Or (2) put

$$u = \frac{1}{3}\cosh^{-1}\left[\frac{G}{2H\sqrt{H}}\right]$$

then the roots are

$$y = -2\sqrt{H}\cosh u, \quad \sqrt{H}\cosh u + \sqrt{-3H}\sinh u, \quad \sqrt{H}\cosh u - \sqrt{-3H}\sinh u$$

Or (3) put

$$u = \frac{1}{3}\sinh^{-1}\left[\frac{G}{2H\sqrt{-H}}\right]$$

Then the roots are

$$y = 2\sqrt{-H}\sinh u, \quad -\sqrt{-H}\sinh u + \sqrt{3H}\cosh u$$
$$-\sqrt{-H}\sinh u - \sqrt{3H}\cosh u$$

SIMULTANEOUS EQUATIONS. Given n independent equations in n unknowns, these n equations usually fix one or more values for each of the n unknowns. To solve such a set of simultaneous equations in x, y, and z, say, solve each of the three equations for x in terms of y and z. Equating these three values for x gives two equations in y and z. Solving each of these two equations for y in terms of z and equating these two values of y gives a single equation in z. The solution of this last equation then gives the value of z. Then substitute this value of z in either of the equations in y and z, and solve for y. Then substitute these values of y and z in any one of the original equations and solve for x.

DETERMINANTS. In the case of linear simultaneous equations (i.e., when x, y, and z occur only to the first power), the equations may be solved by determinants. This method is a considerable time-saver when the number of unknowns is greater than three, but when the number of unknowns is three or less the straight substitution method is preferable.

The determinant of a set of simultaneous equations is formed by writing the equations one below the other with the same unknown in the same relative position in each. The block of numbers forming the coefficients of the unknowns is called the determinant. For example, the determinant of the equations

$$\begin{aligned} w + x + y + z &= 6 \\ w + y + 3z &= 4 \\ w + 2x + 3y &= 1 \\ w + 3x + z &= 3 \end{aligned}$$

is

$$D = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 3 \\ 1 & 2 & 3 & 0 \\ 1 & 3 & 0 & 1 \end{vmatrix}$$

The values of any one of the unknowns, say y, is found by writing a second determinant, D_y, exactly like the determinant D, except that the constants forming the right-hand members of these equations are substituted for the coefficients of y in the determinant, that is

$$D_y = \begin{vmatrix} 1 & 1 & 6 & 1 \\ 1 & 0 & 4 & 3 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{vmatrix}$$

Then

$$y = \frac{D_y}{D}$$

and similarly for the other unknowns.

The value of any determinant is found by making use of the following rules:

(1) If a determinant has two equal rows or columns, it is equal to zero.

(2) To any row or column one may add or subtract any number of times any other row or column without altering the value of the determinant.

(3) To multiply any row or column by a number is the same as multiplying the determinant by that number.

(4) If all the terms in a row or column except one are zero, the determinant reduces to one of a lower order which may be obtained by striking out the row and column which intersect at the element of the row or column which is not zero, and multiplying the whole by that element, changing the sign of this element, however, if it is removed by an odd number of elements from the principal diagonal. The principal diagonal is the line of elements beginning at the upper left-hand corner and ending at the lower right-hand corner. Thus,

$$\begin{vmatrix} 1 & 2 & 8 & 5 \\ 3 & 4 & 6 & 9 \\ 0 & 3 & 0 & 0 \\ 6 & 7 & 4 & 3 \end{vmatrix} = -3 \begin{vmatrix} 1 & 8 & 5 \\ 3 & 6 & 9 \\ 6 & 4 & 3 \end{vmatrix}$$

the principal diagonal being that with the figures 1, 4, 0, and 3. It is immaterial whether the distance from the diagonal is counted along a row or a column.

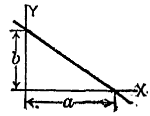
(5) The value of a determinant of the second order is

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1$$

The reduction of determinants is effected by altering the terms according to the above rules until a row or column is obtained in which all terms but one are zero. This enables a reduction of order to be effected in accordance with rule 4. Reductions are continued until one of the second order is obtained.

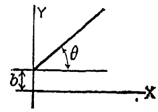
EQUATIONS OF COMMON CURVES. **Straight Line.**

$$\frac{x}{a} + \frac{y}{b} = 1$$



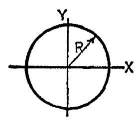
or

$$y = x \tan \theta + b.$$



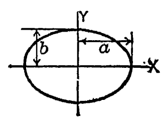
Circle.

$$x^2 + y^2 = R^2$$



Ellipse.

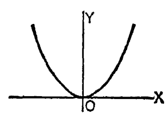
$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$



Parabola (Vertical).

$$y = kx^2$$

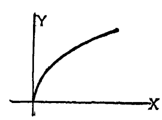
where k is a constant.



Parabola (Horizontal).

$$y = k\sqrt{x}$$

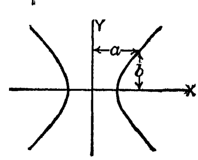
where k is a constant.



Hyperbola.

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \text{ (Horizontal)}$$

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = -1 \text{ (Vertical)}$$



Rectangular or Equilateral Hyperbola.

$$y = \frac{k}{x}$$

where k is a constant.

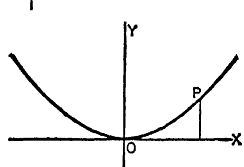
Catenary.

$$y = \frac{1}{k} \cosh kx - 1$$

where k is a constant. The length of arc from O to P is

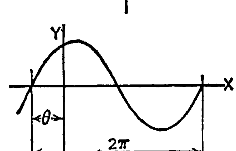
$$= \frac{1}{k} \sinh (kx)$$

See tables of hyperbolic functions.



Sinusoid.

$$y = A \sin (ax + \theta).$$



·2. COMPLEX QUANTITIES

The square root of a negative quantity is called an "imaginary" quantity, or a pure imaginary. A quantity consisting of the sum or difference of a real quantity and an imaginary quantity is called a "complex" quantity. All the rules of ordinary algebra apply to pure imaginaries and complex quantities. The square root of minus one is called the imaginary unit and is usually represented by the symbol j (writers on pure mathematics use the symbol i), that is,

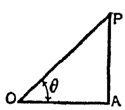
$$j = \sqrt{-1}$$

Any complex quantity may then be written

$$a + jb$$

where a and b are both real quantities.

GEOMETRICAL REPRESENTATION OF A COMPLEX QUANTITY. A positive real quantity may be represented by a line drawn on a plane in a given direction; a negative real quantity may be represented by a line drawn in the opposite direction. Multiplying a quantity by -1 then reverses its direction. Also, since multiplying a real quantity by $\sqrt{-1}$ twice is equivalent to multiplying it by -1, the operation of multiplying once by $\sqrt{-1}$ may be represented by turning the line representing the quantity through 90° in the positive direction of rotation. The positive direction of rotation is taken as the opposite direction to that in which the hands of a clock move. Hence, a complex quantity $a + jb$ may be represented by the line OP in the figure, where $OA = a$ and $AP = b$. The complex quantity $a + jb$ is then completely specified by a line of length $\sqrt{a^2 + b^2}$ making an angle θ, with the axis of reference OX where $\tan\theta = \frac{b}{a}$. The length $M = \sqrt{a^2 + b^2}$ is called the magnitude of the complex quantity, and the angle $\theta = \tan^{-1}\frac{b}{a}$ is called its angle. From the figure it is evident that the complex quantity $a + jb$ may also be written



$$a + jb = M(\cos\theta + j\sin\theta)$$

Expanding $\cos\theta$ and $\sin\theta$ into series (see Series, Article 9) and adding, the resultant series obtained is the series for $e^{j\theta}$; hence

$$a + jb = Me^{j\theta} \tag{1}$$

From the above definitions and equation (1) it is evident that complex numbers possess the following properties:

ADDITION OF TWO COMPLEX QUANTITIES.

$$(a + jb) + (a_1 + jb_1) = (a + a_1) + j(b + b_1)$$

SUBTRACTION OF TWO COMPLEX QUANTITIES.

$$(a + jb) - (a_1 + jb_1) = (a - a_1) + j(b - b_1)$$

MULTIPLICATION OF A COMPLEX QUANTITY BY A COMPLEX NUMBER.

$$(a + jb)(a_1 + jb_1) = aa_1 - bb_1 + j(ab_1 + a_1 b)$$

or, putting $\quad a + jb = Me^{j\theta}$ and $a_1 + jb_1 = M_1 e^{j\theta_1}$

where $\quad M = \sqrt{a^2 + b^2}, \qquad M_1 = \sqrt{a_1{}^2 + b_1{}^2}$

$$\tan\theta = \frac{b}{a}$$

and

$$\tan\theta_1 = \frac{b_1}{a_1}$$

we have $\quad (a + jb)(a_1 + jb_1) = Me^{j\theta} M_1 e^{j\theta_1} = MM_1 e^{j(\theta+\theta_1)}$

Hence the product of two complex quantities is in general a complex quantity which has a magnitude equal to the product of the magnitudes of the two quantities and an angle equal to the sum of the angles of the two quantities.

DIVISION OF A COMPLEX QUANTITY BY A COMPLEX NUMBER.

$$\frac{a + jb}{a_1 + jb_1} = \frac{(a + jb)(a_1 - jb_1)}{(a_1 + jb_1)(a_1 - jb_1)} = \frac{aa_1 + bb_1 - j(ab_1 - a_1 b)}{a_1{}^2 + b_1{}^2}$$

or

$$\frac{a + jb}{a_1 + jb_1} = \frac{Me^{j\theta}}{M_1 e^{j\theta_1}} = \frac{M}{M_1} e^{j(\theta-\theta_1)}$$

Hence the quotient of two complex quantities is in general a complex quantity which has a magnitude equal to the quotient of the magnitudes of the two quantities and an angle equal to the difference of the angles of the two quantities.

SQUARE ROOT OF A COMPLEX QUANTITY.

$$\sqrt{a + jb} = \pm \left[\sqrt{\frac{\sqrt{a^2 + b^2} + a}{2}} + j \sqrt{\frac{\sqrt{a^2 + b^2} - a}{2}} \right]$$

$$\sqrt{a - jb} = \pm \left[\sqrt{\frac{\sqrt{a^2 + b^2} + a}{2}} - j \sqrt{\frac{\sqrt{a^2 + b^2} - a}{2}} \right]$$

and in general

$$\sqrt[N]{a + jb} = M^{1/N} \epsilon^{j\theta/N}$$

Hence the nth root of a complex quantity is, in general, a complex quantity which has a magnitude equal to the nth root of the magnitude M of the quantity and an angle equal one-nth of the angle of the quantity.

EQUATIONS CONTAINING COMPLEX QUANTITIES. Since a real quantity cannot be equal to an imaginary quantity it follows that any equation of the form

$$A + jB = A_1 + jB_1$$

where A, B, A_1, and B_1 are all real quantities (which may, however, consist of any number of terms), is equivalent to the two equations

$$A = A_1$$

and

$$B = B_1$$

Also, if one member of an equation reduces to the form $A + jB$, then the other member of this equation must likewise contain an equal real and an equal imaginary part.

See K. S. Johnson, *Transmission Circuits for Telephonic Communication*, D. Van Nostrand.

3. TRIGONOMETRIC FORMULAS

The trigonometric functions of an angle are the ratios to one another of the various sides of a right triangle having the given angle as one of its angles. Referring to Fig. 1, let B, P, and H be the three sides of a triangle. Then the trigonometric functions of the angle x are

sine of x, abbreviated	$\sin x = \frac{P}{H}$;	cotangent of x, abbreviated	$\cot x = \frac{B}{P}$
cosine of x, abbreviated	$\cos x = \frac{B}{H}$;	secant of x, abbreviated	$\sec x = \frac{H}{B}$
tangent of x, abbreviated	$\tan x = \frac{P}{B}$;	cosecant of x, abbreviated	$\csc x = \frac{H}{P}$

When B, P, and H are limited to the three sides of a right triangle, the above definitions are directly applicable only to angles lying between 0 and 90°. The definitions, however, may be extended by considering the point A (Fig. 2) as describing a circle of radius OA

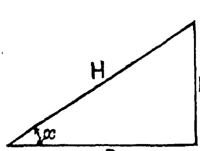


FIG. 1

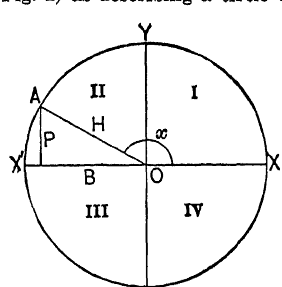


FIG. 2

with the center at O. Let XX' be the horizontal diameter and YY' the vertical diameter of this circle, and call P the perpendicular distance from A to the line XX' and B the horizontal distance from A to YY'. P is to be considered positive when A lies above XX', negative when below. B is considered positive when A is to the right of YY' and negative when to the left. The four quarters of the circle are called quadrants, and are designated as the first, second, third, and fourth quadrants as indicated. The angle is said to lie in the quadrant in which the point A lies. In Fig. 2 the angle x is in the second quadrant.

Algebraic Signs of the Functions

	Sine	Cosine	Tangent
Angle in first quadrant	+	+	+
Angle in second quadrant	+	−	−
Angle in third quadrant	−	−	+
Angle in fourth quadrant	−	+	−

Period. From the above definitions it is evident that adding 2π radians or 360° to an angle does not change the value of any of its functions; that is, these functions repeat themselves every time the angle increases by the 2π radians or 360°. They are therefore said to have a period equal to 2π radians or 360°.

Functions of Angles in Any Quadrant in Terms of Angles in First Quadrant.

$$\begin{aligned}
\sin(-x) &= -\sin x & \sin(90+x) &= \cos x\\
\cos(-x) &= \cos x & \cos(90+x) &= -\sin x\\
\tan(-x) &= -\tan x & \tan(90+x) &= -\cot x\\
\sin(180-x) &= \sin x & \sin(180+x) &= -\sin x\\
\cos(180-x) &= -\cos x & \cos(180+x) &= -\cos x\\
\tan(180-x) &= -\tan x & \tan(180+x) &= \tan x\\
\sin(270-x) &= -\cos x & \sin(270+x) &= -\cos x\\
\cos(270-x) &= -\sin x & \cos(270+x) &= \sin x\\
\tan(270-x) &= \cot x & \tan(270+x) &= -\cot x
\end{aligned}$$

Anti-functions. If $a = \sin x$, then x is the angle whose sine is a; this may be expressed symbolically $x = \sin^{-1} a$, which is read "x equals the angle whose sine is a." The angle x is also called the "anti-sine" or the "inverse sine" of a. Similar notation is used for the other functions; for example, $x = \cos^{-1} b$ is used to express the relation that x is the angle whose cosine is b. At least two "anti-functions" must be known to completely determine the quadrant in which an angle lies; for example, if $x = \sin^{-1} 0.5$ then x may be either 30° or 150°, but if we also have $x = \cos^{-1} 0.866$, then x must equal 30°, while if $x = \cos^{-1}(-0.866)$, then x must equal 150°.

Anti-functions may be taken from the Trigonometric Tables by finding the angle in the margin corresponding to the function in the table.

Example. $\sin^{-1} 0.319 = 18.6°$ or $180° - 18.6° = 161.4°$.

Versine. The expression $(1 - \cos x)$ is called the "versine" of x.

Relations among Functions of the Same Angle.

$$\begin{aligned}
\tan x &= \frac{\sin x}{\cos x} = \frac{1}{\cot x} & \sin^2 x + \cos^2 x &= 1\\
\sec x &= \frac{1}{\cos x} & 1 + \tan^2 x &= \frac{1}{\cos^2 x}\\
\csc x &= \frac{1}{\sin x} & 1 + \cot^2 x &= \frac{1}{\sin^2 x}\\
\sin(90-x) &= \cos x & \sin(-x) &= -\sin x\\
\cos(90-x) &= \sin x & \cos(-x) &= \cos x\\
\tan(90-x) &= \cot x & \tan(-x) &= -\tan x
\end{aligned}$$

Sum and Difference of Two Angles.

$$\begin{aligned}
\sin(x+y) &= \sin x \cos y + \cos x \sin y\\
\cos(x+y) &= \cos x \cos y - \sin x \sin y\\
\tan(x+y) &= \frac{\tan x + \tan y}{1 - \tan x \tan y}\\
\sin(x-y) &= \sin x \cos y - \cos x \sin y\\
\cos(x-y) &= \cos x \cos y + \sin x \sin y\\
\tan(x-y) &= \frac{\tan x - \tan y}{1 + \tan x \tan y}
\end{aligned}$$

Product of the Functions of Two Angles.

$$\sin x \sin y = \tfrac{1}{2}[\cos(x-y) - \cos(x+y)]$$
$$\sin x \cos y = \tfrac{1}{2}[\sin(x+y) + \sin(x-y)]$$
$$\cos x \sin y = \tfrac{1}{2}[\sin(x+y) - \sin(x-y)]$$
$$\cos x \cos y = \tfrac{1}{2}[\cos(x+y) + \cos(x-y)]$$

Functions of Twice an Angle.

$$\sin 2x = 2 \sin x \cos x \qquad \cos 2x = \cos^2 x - \sin^2 x = 2\cos^2 x - 1$$
$$\tan 2x = \frac{2 \tan x}{1 - \tan^2 x}$$

Functions of Half an Angle.

$$\sin\frac{x}{2} = \sqrt{\frac{1-\cos x}{2}} \qquad \cos\frac{x}{2} = \sqrt{\frac{1+\cos x}{2}} \qquad \tan\frac{x}{2} = \sqrt{\frac{1-\cos x}{1+\cos x}}$$

Functions of Three Times an Angle.

$$\sin 3x = 3\sin x - 4\sin^3 x \qquad \cos 3x = 4\cos^3 x - 3\cos x$$
$$\tan 3x = \frac{3\tan x - \tan^3 x}{1 - 3\tan^2 x}$$

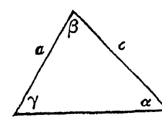


FIG. 3

TRIGONOMETRY. Any triangle is completely defined when (1) two sides and the included angle are known, (2) one side and two angles are known, (3) three sides are known. Let the sides and angles of a triangle be designated as in Fig. 3.

1. Given two sides a and b, and the included angle γ. Then

$$c = \sqrt{a^2 + b^2 - 2ab\cos\gamma}$$
$$\sin\alpha = \frac{a}{c}\sin\gamma$$
$$\beta = 180 - \alpha - \gamma$$

2. Given the side a and the two angles β and γ. Then

$$\alpha = 180 - \beta - \gamma$$
$$b = a\frac{\sin\beta}{\sin\alpha}$$
$$c = a\frac{\sin\gamma}{\sin\alpha}$$

3. Given the three sides a, b, and c. Put

$$s = \tfrac{1}{2}(a + b + c)$$

Then

$$\sin\alpha = \frac{2}{bc}\sqrt{s(s-a)(s-b)(s-c)}$$
$$\sin\beta = \frac{b}{a}\sin\alpha$$
$$\gamma = 180 - \alpha - \beta$$

Relations between Sides and Angles. The following relations between the sides and angles of a triangle are sometimes useful:

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}$$
$$\cos\alpha = \frac{b^2 + c^2 - a^2}{2bc}$$
$$\sin\frac{\alpha}{2} = \sqrt{\frac{(s-b)(s-c)}{bc}}$$
$$\cos\frac{\alpha}{2} = \sqrt{\frac{s(s-a)}{bc}}$$

and similar relations for the other two angles.

4. EXPONENTIAL AND HYPERBOLIC FORMULAS

When the relation between any variable y and another variable x is such that x occurs as an exponent of one or more terms, y is said to be an exponential function of x. Of particular importance in connection with electric circuits are the exponential functions e^x and e^{-x}, where e is the base of the natural logarithms. Since x is the natural logarithm of e^x, the value of e^x can be obtained from the table of common logarithms as shown at the beginning of that table. In addition the values of e^x and e^{-x} are given in a separate table.

Hyperbolic functions are an extension of the trigonometric functions to those cases where the use of the latter gives rise to imaginary or complex angles. From the relations

$$\cos x = \frac{e^{jx} + e^{-jx}}{2}$$

$$\sin x = \frac{e^{jx} - e^{-jx}}{2j}$$

where $j = \sqrt{-1}$, it follows that, putting $x = jz$:

$$\cos jz = \frac{e^z + e^{-z}}{2} \tag{1}$$

$$-j \sin jz = \frac{e^z - e^{-z}}{2} \tag{2}$$

Expressions (1) and (2) are both real quantities when z is real, that is, when the angle jz is imaginary. The first expression is called the hyperbolic cosine of z, abbreviated and pronounced "cosh"; the second expression is called the hyperbolic sine of z, abbreviated sinh and pronounced "shin." Hence, using x for the variable,

$$\sinh x = \frac{e^x - e^{-x}}{2}$$

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

The hyperbolic tangent, cotangent, secant, and cosecant are defined as follows:

$$\tanh x = \frac{\sinh x}{\cosh x}$$

$$\coth x = \frac{\cosh x}{\sinh x}$$

$$\operatorname{sech} x = \frac{1}{\cosh x}$$

$$\operatorname{csch} x = \frac{1}{\sinh x}$$

The hyperbolic angle x is a number analogous to radians in circular measure; it is never expressed in degrees.

Adding 2π to an angle does not change the value of the trigonometric functions; they are therefore said to have a period equal to 2π radians. Hyperbolic functions, however, have no true period, but adding $2\pi j$ to the hyperbolic angle does not change the values of the functions; hence these functions have an imaginary period, $2\pi j$.

For the value of the hyperbolic functions see tables of exponential and hyperbolic functions, Article 13.

Approximate Formulas. Note that, for x less than 0.1,

$$\sinh x = x \text{ with an error of less than 0.2 per cent}$$

$$\cosh x = 1 + \frac{x^2}{2} \text{ with an error of less than 0.09 per cent}$$

For x greater than 6,

$$\sinh x = \cosh x = \frac{e^x}{2} = \frac{1}{2} \log_{10}^{-1} (0.43429x)$$

with an error of less than 0.01 per cent.

Anti-functions. If $a = \sinh x$, then x is the angle whose hyperbolic sine is a; this may be expressed symbolically

$$x = \sinh^{-1} a$$

which is read "x equals the angle whose hyperbolic sine is a." The angle x is also called the "anti-hyperbolic sine" or the "inverse hyperbolic sine" of a. Similarly for the other hyperbolic functions. The following relations exist between the anti-hyperbolic functions and the natural logarithms:

$$\sinh^{-1} x = \log (x + \sqrt{x^2 + 1})$$
$$\cosh^{-1} x = \log (x + \sqrt{x^2 - 1})$$
$$\tanh^{-1} x = \frac{1}{2} \log \left(\frac{1 + x}{1 - x}\right)$$

Relations among Functions of the Same Angle.

$$\cosh^2 x - \sinh^2 x = 1$$
$$1 - \tanh^2 x = \frac{1}{\cosh^2 x}$$
$$\coth^2 x - 1 = \frac{1}{\sinh^2 x}$$
$$\sinh (-x) = - \sinh x$$
$$\cosh (-x) = \cosh x$$
$$\tanh (-x) = - \tanh x$$

See also the definitions given above.

Sum and Difference of Two Angles.

$$\sinh (x + y) = \sinh x \cosh y + \cosh x \sinh y$$
$$\cosh (x + y) = \cosh x \cosh y + \sinh x \sinh y$$
$$\tanh (x + y) = \frac{\tanh x + \tanh y}{1 + \tanh x \tanh y}$$
$$\sinh (x - y) = \sinh x \cosh y - \cosh x \sinh y$$
$$\cosh (x - y) = \cosh x \cosh y - \sinh x \sinh y$$
$$\tanh (x - y) = \frac{\tanh x - \tanh y}{1 - \tanh x \tanh y}$$

Product of the Functions of Two Angles.

$$\sinh x \sinh y = {}^1/_2 [\cosh (x + y) - \cosh (x - y)]$$
$$\sinh x \cosh y = {}^1/_2 [\sinh (x + y) + \sinh (x - y)]$$
$$\cosh x \sinh y = {}^1/_2 [\sinh (x + y) - \sinh (x - y)]$$
$$\cosh x \cosh y = {}^1/_2 [\cosh (x + y) + \cosh (x - y)]$$

Functions of Twice an Angle.

$$\sinh 2x = 2 \sinh x \cosh x$$
$$\cosh 2x = \sinh^2 x + \cosh^2 x = 2 \sinh^2 x + 1 = 2 \cosh^2 x - 1$$
$$\tanh 2x = \frac{2 \tanh x}{1 + \tanh^2 x}$$

Functions of Half an Angle.

$$\sinh \frac{x}{2} = \sqrt{\frac{\cosh x - 1}{2}}$$
$$\cosh \frac{x}{2} = \sqrt{\frac{\cosh x + 1}{2}}$$
$$\tanh \frac{x}{2} = \sqrt{\frac{\cosh x - 1}{\cosh x + 1}}$$

Functions of Three Times an Angle.

$$\sinh 3x = 3 \sinh x + 4 \sinh^3 x$$
$$\cosh 3x = 4 \cosh^3 x - 3 \cosh x$$
$$\tanh 3x = \frac{3 \tanh x + \tanh^3 x}{1 + 3 \tanh^2 x}$$

Relations between Hyperbolic and Trigonometric Functions.

$$\sinh (jx) = j \sin x \qquad \sin (jx) = j \sinh x$$
$$\cosh (jx) = \cos x \qquad \cos (jx) = \cosh x$$
$$\tanh (jx) = j \tan x \qquad \tan (jx) = j \tanh x$$
$$\sinh^{-1} jx = j \sin^{-1} x \qquad \sin^{-1} jx = j \sinh^{-1} x$$
$$\tanh^{-1} jx = j \tan^{-1} x \qquad \tan^{-1} jx = j \tanh^{-1} x$$

$$\cosh^{-1} jx = j \cos^{-1} jx = \log (x + \sqrt{1 + x^2}) - j \frac{\pi}{2}$$

Hyperbolic Functions of a Complex Angle.

$$\sinh (x + jy) = \sinh x \cos y + j \cosh x \sin y = Me^{j\theta}$$

where $$M = \sqrt{\frac{\cosh 2x - \cos 2y}{2}} \quad \text{and} \quad \tan \theta = \frac{\tan y}{\tanh x}$$

$$\cosh (x + jy) = \cosh x \cos y + j \sinh x \sin y = Ne^{j\phi}$$

where $$N = \sqrt{\frac{\cosh 2x + \cos 2y}{2}} \quad \text{and} \quad \tan \phi = \tanh x \cdot \tan y$$

$$\tanh (x + jy) = \frac{\sinh x \cos y + j \cosh x \sin y}{\cosh x \cos y + j \sinh x \sin y} = Pe^{j\psi}$$

where $$P = \sqrt{\frac{\cosh 2x - \cos 2y}{\cosh 2x + \cos 2y}} \quad \text{and} \quad \psi = \tan^{-1} \left[\frac{\sin 2y}{\sinh 2x}\right]$$

$$\tanh^{-1} (Ae^{j\alpha}) = B_1 + jB_2$$

where $$B_1 = \frac{1}{2} \tanh^{-1} \left[\frac{2A \cos \alpha}{1 + A^2}\right] \quad \text{and} \quad B_2 = \frac{1}{2} \tan^{-1} \left[\frac{2A \sin \alpha}{1 - A^2}\right]$$

5. CALCULUS FORMULAS

The formula for the integration by parts is:

$$\int_a^b u\, dv = [uv]_a^b - \int_a^b v\, du$$

The following table is used in the formulas

$$\frac{df(x)}{dx} = f'(x)$$

$$\int f'(x)\, dx = f(x) + C$$

where C is an arbitrary constant.

$f'(x)$	$f(x)$	$f'(x)$	$f(x)$
x^m	$\frac{1}{m+1} x^{m+1}$	$\frac{1}{\cos^2 ax}$	$\frac{1}{a} \tan ax$
$\frac{1}{ax}$	$\frac{1}{a} \log_e x$	$\frac{1}{\sin^2 ax}$	$-\frac{1}{a} \cot ax$
e^{ax}	$\frac{1}{a} e^{ax}$	$\frac{1}{\sqrt{a^2 + bx^2}}$	$\frac{1}{\sqrt{-b}} \sin^{-1} \sqrt{-b}\, \frac{x}{a}$
a^{bx}	$\frac{1}{b \log a} a^{bx}$	$\frac{1}{x\sqrt{x^2 + a}}$	$\frac{1}{\sqrt{-a}} \cos^{-1} \sqrt{-a}\, \frac{1}{x}$
$\cos ax$	$\frac{1}{a} \sin ax$	$\frac{x}{\sqrt{a^2 \pm x^2}}$	$\pm\sqrt{a^2 \pm x^2}$
$\sin ax$	$-\frac{1}{a} \cos ax$	$\frac{x}{\sqrt{x^2 - a^2}}$	$\sqrt{x^2 - a^2}$
$\cosh ax$	$\frac{1}{a} \sinh ax$	$\frac{u \frac{dv}{dx} - v \frac{du}{dx}}{u^2}$	$\frac{v}{u}$
$\sinh ax$	$\frac{1}{a} \cosh ax$	$\log x$	$x \log x - x$
$\tan ax$	$-\frac{1}{a} \log (\cos ax)$	$\sin^2 x$	$-{}^1/_2(\cos x \sin x - x)$
$\tanh ax$	$\frac{1}{a} \log (\cosh ax)$	$\cos^2 x$	${}^1/_2(\sin x \cos x + x)$

MAXIMA AND MINIMA. Let y be any function of a variable x; then y will be a maximum or minimum for any value of x which satisfies

$$\frac{dy}{dx} = 0 \tag{1}$$

provided $\frac{d^2y}{dx^2}$ is not zero. If the second derivative $\frac{d^2y}{dx^2}$ is positive for this value of x, then the corresponding value of y is a minimum; if this second derivative is negative, the corresponding value of y is a maximum.

In case $\frac{d^2y}{dx^2}$ is also zero for the value of x which satisfies (1), the corresponding value of y is not a maximum or minimum unless $\frac{d^3y}{dx^3}$ is also zero and $\frac{d^4y}{dx^4}$ is not zero. When $\frac{d^3y}{dx^3} = 0$, y is a minimum if $\frac{d^4y}{dx^4}$ is positive and a maximum if $\frac{d^4y}{dx^4}$ is negative. In case $\frac{d^4y}{dx^4}$ is also zero, similar relations must hold for the fifth and sixth derivatives, etc.

6. DIFFERENTIAL EQUATIONS

Differential equations of the following forms are met with in the theory of alternating and transient currents.

The following notation is used: $e = 2.7183\cdots$ = base of natural system of logarithms; x, y, z are variables. A, ϕ, γ, and θ are constants of integration or arbitrary constants. Other letters represent known constants.

$$\frac{dy}{dx} = ay \tag{1}$$

Solution:

$$y = Ae^{ax}$$

$$\frac{dy}{dx} + ay = 0 \tag{2}$$

Solution:

$$y = Ae^{-ax}$$

$$\frac{dy}{dx} + ay = b \tag{3}$$

Solution:

$$y = \frac{b}{a}[1 - Ae^{-ax}]$$

$$\frac{d^2y}{dx^2} = -a^2y \tag{4}$$

Solution:

$$y = A \sin (ax + \phi)$$

$$\frac{d^2y}{dx^2} = a^2y \tag{5}$$

Solution:

$$y = A \sinh (ax + \phi)$$

$$\frac{d^2y}{dx^2} + 2u\frac{dy}{dx} + (u^2 + a^2)y = 0 \tag{6}$$

Solution:

Case I. a^2 positive: $y = Ae^{-ux} \sin (ax + \phi)$
Case II. a^2 negative: $y = Ae^{-ux} \sinh (ax + \phi)$
Case III. $a^2 = 0$: $y = A(x + \phi)e^{-ux}$

$$\frac{d^2y}{dx^2} + 2u\frac{dy}{dx} + (u^2 + a^2)y = B \sin (\omega x + \theta) \tag{7}$$

The complete solution of this equation consists of the solution of (6) plus the term

$$\left(\frac{B \sin \delta}{2u\omega}\right) \sin (\omega x + \theta - \delta) \tag{a}$$

where

$$\delta = \tan^{-1} \frac{2u\omega}{a^2 + u^2 - \omega^2}$$

For each additional sine term added to the right-hand member of the equation, there will be a corresponding term of the same form as (a) in the solution.

$$\frac{d^n y}{dx^n} + a_{n-1}\frac{d^{n-1}y}{dx^{n-1}} + \cdots a_1\frac{dy}{dx} + a_0 y = B\sin(\omega x + \theta) \tag{8}$$

Solution:

$$y = A_1 e^{m_1 x} + A_2 e^{m_2 x} + \cdots A_n e^{m_n x} + KB\sin(\omega X + \theta + \delta)$$

where m_1, m_2, etc , are the n roots of the equation

$$m^n + a_{n-1}m^{n-1} + \cdots a_1 m + a_0 = 0$$

and K and δ are found by substituting the $KB\sin(\omega x + \theta + \delta)$ by itself in the given differential equation and equating the coefficients of $\sin(\omega x + \theta)$ and $\cos(\omega x + \theta)$ respectively on the two sides of the resulting equation. When the second member of the differential equation is a constant, B, the sine term in the solution becomes simply $\frac{B}{a_0}$.

Note that all the preceding equations are merely special cases of the general equation (8)

$$\frac{d^2y}{dx^2} + 2u\frac{dy}{dx} + (u^2 - q^2)y = \frac{1}{c^2}\frac{d^2y}{dz^2} \tag{9}$$

The complete solution of this equation contains an infinite number of terms of the form

$$y = e^{-(u-s)x}\left[A_1 e^{mz}\sin(\omega x + nz + \phi_1) + A_2 e^{-mz}\sin(\omega x - nz + \phi_2)\right] \tag{a}$$

where A_1, ϕ_1, A_2, ϕ_2, and two of the four constants ω, s, m, and n are integration constants (fixed by the terminal conditions). The values of m and n in terms of ω and s are

$$m = c\sqrt{ab}\cos\frac{\eta + e}{2}$$

$$n = c\sqrt{ab}\sin\frac{\eta + e}{2}$$

where $a = \sqrt{(s+q)^2 + \omega^2}$, $\quad e = \tan^{-1}\left(\frac{\omega}{s+q}\right)$

$b = \sqrt{(s-q)^2 + \omega^2}$, $\quad \eta = \tan^{-1}\left(\frac{\omega}{s-q}\right)$

The values of ω and s in terms of m and n are

$$\omega = \frac{\sqrt{FG}}{c}\cos\frac{\alpha + \beta}{2}$$

$$s = \frac{\sqrt{FG}}{c}\sin\frac{\alpha + \beta}{2}$$

where $F = \sqrt{(n+cq)^2 + m^2}$, $\quad \alpha = \tan^{-1}\left(\frac{m}{n+cq}\right)$

$G = \sqrt{(n-cq)^2 + m^2}$, $\quad \beta = \tan^{-1}\left(\frac{m}{n-cq}\right)$

The solution of eq. (9) may also be written as a series of terms of the form

$$y = Me^{-(u-s)x}\sin(\omega x + \phi + \mu) \tag{b}$$

where

$$M = \frac{A}{\sqrt{2}}\sqrt{\cosh 2(mz + \gamma) + \cos 2(nz + \theta)}$$

$$\tan\mu = \tanh(mz + \gamma)\tan(nz + \theta)$$

where A, ϕ, γ, and θ are integration constants, and the relations between the other constants ω, s, m, and n are the same as above.

In the special case when $q = 0$, the solution of eq. (9) is

$$y = e^{-ux}\left[f_1(\omega x + nz) + f_2(\omega x - nz)\right] \tag{c}$$

where f_1 and f_2 are any two arbitrary functions and ω and n are connected by the relation

$$\frac{\omega}{n} = \frac{1}{c}$$

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + \left(1 - \frac{n^2}{x^2}\right) y = 0 \tag{10}$$

This is known as Bessel's equation of order n. $J_n(x)$, Bessel's function of the first kind of order n, is a particular solution of this equation. It may be computed from the infinite series:

$$J_n(x) = \frac{x^n}{2^n\Gamma(n+1)}\left[1 - \frac{x^2}{2^2(n+1)} + \frac{x^4}{2^4 2!(n+1)(n+2)} - \frac{x^6}{2^6 3!(n+1)(n+2)(n+3)} + \cdots\right] \tag{a}$$

where $\Gamma(n + 1)$ is the gamma function which reduces to unity for $n = 0$ and to $n!$ for n equal to any positive integer. In general, the function $J_n(x)$ is an oscillatory function of x having the value zero for $x = 0$, except for the case where $n = 0$. For values of n larger than 1, the slope of $J_n(x)$ is zero for $x = 0$ and the first maximum and the first zero occurs at successively higher values of x as n takes on larger values. For small values of n, the values of x for which $J_n(x)$ is a maximum or zero can be gotten from tables of Bessel functions. For large values of n, the first maximum, that is, the smallest x for which $J_n'(x) = 0$, is given by

$$n + 0.809\sqrt[3]{n} \tag{b}$$

with an error not larger than $1/\sqrt[3]{n}$, and the first zero, that is, the smallest x for which $J_n(x) = 0$, is given by

$$n + 1.856\sqrt[3]{n} \tag{c}$$

again with an error of the order of $1/\sqrt[3]{n}$.

For integral values of n greater than zero,

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x) \tag{d}$$

which permits one to compute Bessel functions for successively higher order from tables of $J_0(x)$ and $J_1(x)$.

When n is an integer,

$$J_{-n}(x) = (-1)^n J_n(x) \tag{e}$$

BIBLIOGRAPHY

Watson, G. N., *A Treatise on the Theory of Bessel Functions*, 2nd Ed., New York, The Macmillan Company (1944).
Article 14 of this section.

7. ERRORS OF OBSERVATION

When a quantity is measured with all possible accuracy many times in succession, the numbers expressing the results are found to differ by amounts which, although generally small, are occasionally considerable in comparison with the quantity measured. Though these differences may be decreased by improved methods, better instruments, or greater skill, they can never be entirely removed. They are known as the errors of observation. The following formulas, which are derived from the theory of least squares, apply to such errors and not to errors which can be eliminated by correcting mistakes of the observer or defects of instruments or methods of observation. That is, they apply only to errors which may be either positive or negative, the chance of a positive error occurring being exactly the same as the chance of a negative error occurring.

WEIGHTED OBSERVATIONS. Sometimes, in spite of the care with which observations are taken, there are reasons for believing that some observations are better than others. In this case the observations are given different "weights" or numbers expressing their relative practical worth. A weighted observation is an observation multiplied by its weight.

PROBABLE VALUE OF SEVERAL OBSERVATIONS. The most probable value of a quantity which is observed directly several times with equal care is the arithmetical mean of the measurements.

The most probable value of a quantity which is observed directly several times, but the observations of which have different weights, is equal to the sum of the weighted observations divided by the sum of the weights.

PROBABLE ERROR OF ANY ONE OF SEVERAL OBSERVATIONS. The probable error or dispersion of a number of direct observations made with equal care is given by the following formula:

$$r = 0.6745\sqrt{\frac{\Sigma v^2}{n-1}}$$

where n = number of observations.
r = probable error of a single observation.
v = residual found by subtracting the arithmetical mean from each measurement.

The probable error of each of a number of direct observations, where the observations have different weight, is found by the following formula, in which p represents the per unit weight of an observation.

$$r_1 = 0.6745\sqrt{\frac{\Sigma pv^2}{n-1}}$$

PROBABLE ERROR OF THE ARITHMETICAL MEAN. If

r = probable error of a single observation,
n = number of observations,
r_0 = probable error of the arithmetical mean,

$$r_0 = \frac{r}{\sqrt{n}} \text{ for observations of equal weight}$$

or

$$r_0 = \frac{r_1}{\sqrt{\Sigma p}} \text{ for unequal weight}$$

It should be noted that the probable error of the mean decreases inversely as the square root of the number of observations.

PROBABLE ERROR IN A RESULT CALCULATED FROM THE MEANS OF SEVERAL OBSERVED QUANTITIES. Let Z = a sum or difference of several independent quantities.

Let r_1, r_2, r_3, etc., be the probable errors in these quantities. Then the probable error of Z is equal to

$$\sqrt{r_1^2 + r_2^2 + r_3^2 + \text{etc.}}$$

Let $Z = Az$, where z is an observed quantity, and A, a known number. Let r be the probable error in z. Then the probable error in Z is Ar.

Let Z be the product of two independently observed quantities z_1 and z_2 whose probable errors are r_1 and r_2 respectively. Then the error in Z is equal to

$$\sqrt{z_1^2 r_2^2 + z_2^2 r_1^2}$$

Let Z be any function of the independently observed quantities z_1, z_2, z_3, etc., whose probable errors are r_1, r_2, r_3, etc. Then the probable error in Z is equal to

$$\sqrt{\left(\frac{\partial Z}{\partial z_1}\right)^2 r_1^2 + \left(\frac{\partial Z}{\partial z_2}\right)^2 r_2^2 + \left(\frac{\partial Z}{\partial z_3}\right)^3 r_3^2 + \text{etc.}}$$

8. APPROXIMATIONS

If a is small

$$(1 \pm a)^m = 1 \pm ma$$

If m is nearly equal to n

$$\sqrt{mn} = \frac{m+n}{2}$$

If θ, expressed in radians, is small compared to a radian

$$\sin\theta = \tan\theta = \theta \text{ radians}$$

9. SERIES

Taylor's series is written

$$f(x+h) = f(x) + \frac{h}{1!}f'(x) + \frac{h^2}{2!}f''(x) + \cdots$$
$$= f(h) + \frac{x}{1!}f'(h) + \frac{x^2}{2!}f''(h) + \cdots$$

where the prime on the function means the derivative with respect to the argument.

The following series are frequently useful.

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$a^x = 1 + x\log a + \frac{(x\log a)^2}{2!} + \frac{(x\log a)^3}{3!} + \cdots$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

$$\cos(x\sin\theta) = J_0(x) + 2\{J_2(x)\cos 2\theta + J_4(x)\cos 4\theta + \cdots$$

where $J_n(x)$ is Bessel's function of order n,

$$\sin(x\sin\theta) = 2\{J_1(x)\sin\theta + J_3(x)\sin 3\theta + \cdots$$

$$\cos(x\cos\theta) = J_0(x) - 2J_2(x)\cos 2\theta + 2J_4(x)\cos 4\theta + \cdots$$

$$\sin(x\cos\theta) = 2J_1(x)\cos\theta - 2J_3(x)\cos 3\theta + 2J_5(x)\cos 5\theta + \cdots$$

$$\begin{aligned}\sin(A + x\sin\theta) = {} & J_0(x)\sin A + J_1(x)[\sin(A+\theta) - \sin(A-\theta)] \\ & + J_2(x)[\sin(A+2\theta) + \sin(A-2\theta)] \\ & + J_3(x)[\sin(A+3\theta) - \sin(A-3\theta)] \\ & + J_4(x)[\sin(A+4\theta) + \sin(A-4\theta)] + \cdots\end{aligned}$$

10. MENSURATION

The term mensuration is used in this article to include the relations between the areas and volumes of geometric figures and their linear dimensions.

Triangle.

$$\text{Area} = {}^1/_2\,(\text{Base}) \times (\text{Perpendicular height})$$
$$= \sqrt{s(s-a)(s-b)(s-c)}$$

where a, b, and c are the lengths of the three sides respectively, and $s = {}^1/_2\,(a+b+c)$.

Trapezoid.

$$\text{Area} = \left(\frac{a+b}{2}\right)d$$

where a and b are the lengths of the parallel sides respectively, and d their distance apart.

Parallelogram.

$$\text{Area} = (\text{Base}) \times (\text{Perpendicular height})$$

Parabola.

$$\text{Area} = {}^2/_3\,(\text{Area of circumscribing rectangle})$$

Cycloid.

$$\text{Area} = {}^3/_4\,\pi x(\text{Altitude})^2$$

the altitude being the diameter of the rolling circle.

Circle.

$$\text{Circumference} = 2\pi r = \pi d$$

$$\text{Area} = \pi r^2 = \frac{\pi}{4} d^2$$

where r is the radius and d the diameter.

$$\text{Area of segment} = \frac{r^2}{2} (\theta - \sin \theta)$$

where θ is the angle in radians (see Angles) subtended by the arc of the segment. If n is the height of the segment, measured along the radius perpendicular to the chord,

$$\text{Area of segment} = \pi r^2 M - A(r - n)$$

where $$A = \sqrt{n(2r - n)} \quad \text{and} \quad M = \frac{1}{180} \sin^{-1} \left(\frac{A}{r}\right)$$

Ellipse.

$$\text{Area} = \pi ab$$

where a and b are the principal semi-axes.

Prism with Parallel Sides and Parallel Ends.

$$\text{Volume} = (\text{Area of end}) \times (\text{Perpendicular distance between ends})$$

Right Circular Cylinder.

$$\text{Volume} = \frac{\pi}{4} d^2 l$$

where d is the diameter and l the length.

$$\text{Total surface of right cylinder} = \pi d(l + {}^1\!/_2 d)$$

Right Circular Cone.

$$\text{Volume} = {}^1\!/_3 \,(\text{Area of base}) \times (\text{Height})$$

$$= {}^1\!/_3 \,(\text{Volume of circumscribing cylinder})$$

where r is the radius of base and h the height of the cone.

$$\text{Area of curved surface of a right circular cone} = \pi r \sqrt{h^2 + r^2}$$

Right Pyramid.

$$\text{Volume} = {}^1\!/_3 \,(\text{Area of base}) \times (\text{Height}).$$

$$\text{Volume of frustum of pyramid} = {}^1\!/_3 \,(\text{Height}) \,(A + \sqrt{aA})$$

where A and a are the areas of the ends respectively.

Sphere.

$$r = \text{radius}$$

$$\text{Area of surface} = 4\pi r^2 = {}^2\!/_3 \,(\text{total area of circumscribing cylinder})$$

Area of the surface of a zone of a sphere = area of zone of the same height as this zone projected on to a cylinder.

$$\text{Volume} = {}^4\!/_3 \,\pi r^3 = {}^2\!/_3 \,(\text{volume of circumscribing cylinder})$$

Volume of a frustum of a sphere $= \pi r^2(k \pm h) - \frac{\pi}{3} (k^3 \pm h^3)$, where k is the distance of its outer face from center and h the distance of its inner face from the center, the negative signs in the brackets to be used if both faces are on the same side of the center and the positive signs if on opposite sides of the center.

Ellipsoid.

$$\text{Volume} = {}^4\!/_3 \,\pi abc$$

where a, b, and c are the three principal semi-axes, respectively.

Paraboloid. Volume of a paraboloid of revolution equals one-half that of the circumscribing cylinder.

MATHEMATICAL TABLES AND CHARTS

11. COMMON AND NATURAL LOGARITHMS OF NUMBERS

The common logarithm of a number is the index of the power to which the base 10 must be raised in order to equal the number.

The common logarithm of every positive number not an integral power of 10 consists of an *integral* and a *decimal part.* The integral part or whole number is called the *characteristic* and may be either *positive or negative.* The decimal or fractional part is a *positive* number called the *mantissa* and is the same for all numbers which have the same sequential digits.

The characteristic of the logarithm of any positive number greater than one is positive and is one less than the number of digits before the decimal point.

The characteristic of the logarithm of any positive number less than one is negative and is one more than the number of ciphers immediately after the decimal point.

A negative number or number less than zero has no real logarithm.

EXAMPLES: $\text{Log}_{10}\,25400. = 4.404834$ $\text{Log}_{10}\,0.0254 = \bar{2}.404834$ or $8.404834 - 10$

The two systems of logarithms in general use are the common or Briggsian logarithms, introduced in 1615 by Henry Briggs, a contemporary of John Napier, the inventor of logarithms, and the natural or less appropriately termed Napierian or hyperbolic logarithms, which developed somewhat accidentally from Napier's original work. The latter have a base denoted by e, an irrational number, which is:

$$e = \text{Lim}_{u \to \infty} \left(1 + \frac{1}{u}\right)^u = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots = 2.7182818$$

To obtain the natural logarithm, the common logarithm given below is multiplied by $\log_e 10$ which is 2.302585, or $\log_e N = 2.302585 \log_{10} N$.

N	0	1	2	3	4	5	6	7	8	9
0		**000000**	**301030**	**477121**	**602060**	**698970**	**778151**	**845098**	**903090**	**954243**
1	000000	041393	079181	113943	146128	176091	204120	230449	255273	278754
2	301030	322219	342423	361728	380211	397940	414973	431364	447158	462398
3	477121	491362	505150	518514	531479	544068	556303	568202	579784	591065
4	602060	612784	623249	633468	643453	653213	662758	672098	681241	690196
5	698970	707570	716003	724276	732394	740363	748188	755875	763428	770852
6	778151	785330	792392	799341	806180	812913	819544	826075	832509	838849
7	845098	851258	857332	863323	869232	875061	880814	886491	892095	897627
8	903090	908485	913814	919078	924279	929419	934498	939519	944483	949390
9	954243	959041	963788	968483	973128	977724	982271	986772	991226	995635
10	**000000**	**004321**	**008600**	**012837**	**017033**	**021189**	**025306**	**029384**	**033424**	**037426**
1	041393	045323	049218	053078	056905	060698	064458	068186	071882	075547
2	079181	082785	086360	089905	093422	096910	100371	103804	107210	110590
3	113943	117271	120574	123852	127105	130334	133539	136721	139879	143015
4	146128	149219	152288	155336	158362	161368	164353	167317	170262	173186
5	176091	178977	181844	184691	187521	190332	193125	195900	198657	201397
6	204120	206826	209515	212188	214844	217484	220108	222716	225309	227887
7	230449	232996	235528	238046	240549	243038	245513	247973	250420	252853
8	255273	257679	260071	262451	264818	267172	269513	271842	274158	276462
9	278754	281033	283301	285557	287802	290035	292256	294466	296665	298853
20	**301030**	**303196**	**305351**	**307496**	**309630**	**311754**	**313867**	**315970**	**318063**	**320146**
1	322219	324282	326336	328380	330414	332438	334454	336460	338456	340444
2	342423	344392	346353	348305	350248	352183	354108	356026	357935	359835
3	361728	363612	365488	367356	369216	371068	372912	374748	376577	378398
4	380211	382017	383815	385606	387390	389166	390935	392697	394452	396199
5	397940	399674	401401	403121	404834	406540	408240	409933	411620	413300
6	414973	416641	418301	419956	421604	423246	424882	426511	428135	429752
7	431364	432969	434569	436163	437751	439333	440909	442480	444045	445604
8	447158	448706	450249	451786	453318	454845	456366	457882	459392	460898
9	462398	463893	465383	466868	468347	469822	471292	472756	474216	475671
30	**477121**	**478566**	**480007**	**481443**	**482874**	**484300**	**485721**	**487138**	**488551**	**489958**
1	491362	492760	494155	495544	496930	498311	499687	501059	502427	503791
2	505150	506505	507856	509203	510545	511883	513218	514548	515874	517196
3	518514	519828	521138	522444	523746	525045	526339	527630	528917	530200
4	531479	532754	534026	535294	536558	537819	539076	540329	541579	542825
5	544068	545307	546543	547775	549003	550228	551450	552668	553883	555094

N	0	1	2	3	4	5	6	7	8	9
5	544068	545307	546543	547775	549003	550228	551450	552668	553883	555094
6	556303	557507	558709	559907	561101	562293	563481	564666	565848	567026
7	568202	569374	570543	571709	572872	574031	575188	576341	577492	578639
8	579784	580925	582063	583199	584331	585461	586587	587711	588832	589950
9	591065	592177	593286	594393	595496	596597	597695	598791	599883	600973
40	**602060**	**603144**	**604226**	**605305**	**606381**	**607455**	**608526**	**609594**	**610660**	**611723**
1	612784	613842	614897	615950	617000	618048	619093	620136	621176	622214
2	623249	624232	625312	626340	627366	628389	629410	630428	631444	632457
3	633468	634477	635484	636488	637490	638489	639486	640481	641474	642465
4	643453	644439	645422	646404	647383	648360	649335	650308	651278	652246
5	653213	654177	655138	656098	657056	658011	658965	659916	660865	661713
6	662758	663701	664642	665581	666518	667453	668386	669317	670246	671173
7	672098	673021	673942	674861	675778	676694	677607	678518	679428	680336
8	681241	682145	683047	683947	684845	685742	686636	687529	688420	689309
9	690196	691081	691965	692847	693727	694605	695482	696356	697229	698100
50	**698970**	**699838**	**700704**	**701568**	**702431**	**703291**	**704151**	**705008**	**705864**	**706718**
1	707570	708421	709270	710117	710963	711807	712650	713491	714330	715167
2	716003	716838	717671	718502	719331	720159	720986	721811	722634	723456
3	724276	725095	725912	726727	727541	728354	729165	729974	730782	731589
4	732394	733197	733999	734800	735599	736397	737193	737987	738781	739572
5	740363	741152	741939	742725	743510	744293	745075	745855	746634	747412
6	748188	748963	749736	750508	751279	752048	752816	753583	754348	755112
7	755875	756636	757396	758155	758912	759668	760422	761176	761928	762679
8	763428	764176	764923	765669	766413	767156	767898	768638	769377	770115
9	770852	771587	772322	773055	773786	774517	775246	775974	776701	777427
60	**778151**	**778874**	**779596**	**780317**	**781037**	**781755**	**782473**	**783189**	**783904**	**784617**
1	785330	786041	786751	787460	788168	788875	789581	790285	790988	791691
2	792392	793092	793790	794488	795185	795880	796574	797268	797960	798651
3	799341	800029	800717	801404	802089	802774	803457	804139	804821	805501
4	806180	806858	807535	808211	808886	809560	810233	810904	811575	812245
5	812913	813581	814248	814913	815578	816241	816904	817565	818226	818885
6	819544	820201	820858	821514	822168	822822	823474	824126	824776	825426
7	826075	826723	827369	828015	828660	829304	829947	830589	831230	831870
8	832509	833147	833784	834421	835056	835691	836324	836957	837588	838219
9	838849	839478	840106	840733	841359	841985	842609	843233	843855	844477
70	**845098**	**845718**	**846337**	**846955**	**847573**	**848189**	**848805**	**849419**	**850033**	**850646**
1	851258	851870	852480	853090	853698	854306	854913	855519	856124	856729
2	857332	857935	858537	859138	859739	860338	860937	861534	862131	862728
3	863323	863917	864511	865104	865696	866287	866878	867467	868056	868644
4	869232	869818	870404	870989	871573	872156	872739	873321	873902	874482
5	875061	875640	876218	876795	877371	877947	878522	879096	879669	880242
6	880814	881385	881955	882525	883093	883661	884229	884795	885361	885926
7	886491	887054	887617	888179	888741	889302	889862	890421	890980	891537
8	892095	892651	893207	893762	894316	894870	895423	895975	896526	897077
9	897627	898176	898725	899273	899821	900367	900913	901458	902003	902547
80	**903090**	**903633**	**904174**	**904716**	**905256**	**905796**	**906335**	**906874**	**907411**	**907949**
1	908485	909021	909556	910091	910624	911158	911690	912222	912753	913284
2	913814	914343	914872	915400	915927	916454	916980	917506	918030	918555
3	919078	919601	920123	920645	921166	921686	922206	922725	923244	923762
4	924279	924796	925312	925828	926342	926857	927370	927883	928396	928908
5	929419	929930	930440	930949	931458	931966	932474	932981	933487	933993
6	934498	935003	935507	936011	936514	937016	937518	938019	938520	939020
7	939519	940018	940516	941014	941511	942008	942504	943000	943495	943989
8	944483	944976	945469	945961	946452	946943	947434	947924	948413	948902
9	949390	949878	950365	950851	951338	951823	952308	952792	953276	953760
90	**954243**	**954725**	**955207**	**955688**	**956168**	**956649**	**957128**	**957607**	**958086**	**958564**
1	959041	959518	959995	960471	960946	961421	961895	962369	962843	963316
2	963788	964260	964731	965202	965672	966142	966611	967080	967548	968016
3	968483	968950	969416	969882	970347	970812	971276	971740	972203	972666
4	973128	973590	974051	974512	974972	975432	975891	976350	976808	977266
5	977724	978181	978637	979093	979548	980003	980458	980912	981366	981819
6	982271	982723	983175	983626	984077	984527	984977	985426	985875	986324
7	986772	987219	987666	988113	988559	989005	989450	989895	990339	990783
8	991226	991669	992111	992554	992995	993436	993877	994317	994757	995196
9	995635	996074	996512	996949	997386	997823	998259	998695	999131	999565
100	**000000**	**000434**	**000868**	**001301**	**001734**	**002166**	**002598**	**003029**	**003461**	**003891**

12. TRIGONOMETRIC TABLES

The following tables give the values of sin x, cos x, and tan x for values of x from 0 to 90° in intervals of 0.1 degree. By making use of the periodic character of these functions, the values can be determined from these tables for all values of x to an accuracy of 0.1 degree. (See Trigonometric Formulas.)

If the angle is given in radians multiply the number of radians by $\frac{180}{\pi}$ (57.295) to obtain the number of degrees.

Trigonometric Functions 0.0°–15.9°

Angle in Degrees	Name of Function	Value of Function for Each Tenth of a Degree									
		0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
	sin	0.0000	0.0017	0.0035	0.0052	0.0070	0.0087	0.0105	0.0122	0.0140	0.0157
0	cos	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999
	tan	0.0000	0.0017	0.0035	0.0052	0.0070	0.0087	0.0105	0.0122	0.0140	0.0157
	sin	0.0175	0.0192	0.0209	0.0227	0.0244	0.0262	0.0279	0.0297	0.0314	0.0332
1	cos	0.9998	0.9998	0.9998	0.9997	0.9997	0.9997	0.9996	0.9996	0.9995	0.9995
	tan	0.0175	0.0192	0.0209	0.0227	0.0244	0.0262	0.0279	0.0297	0.0314	0.0332
	sin	0.0349	0.0366	0.0384	0.0401	0.0419	0.0436	0.0454	0.0471	0.0488	0.0506
2	cos	0.9994	0.9993	0.9993	0.9992	0.9991	0.9990	0.9990	0.9989	0.9988	0.9987
	tan	0.0349	0.0367	0.0384	0.0402	0.0419	0.0437	0.0454	0.0472	0.0489	0.0507
	sin	0.0523	0.0541	0.0558	0.0576	0.0593	0.0610	0.0628	0.0645	0.0663	0.0680
3	cos	0.9986	0.9985	0.9984	0.9983	0.9982	0.9981	0.9980	0.9979	0.9978	0.9977
	tan	0.0524	0.0542	0.0559	0.0577	0.0594	0.0612	0.0629	0.0647	0.0664	0.0682
	sin	0.0698	0.0715	0.0732	0.0750	0.0767	0.0785	0.0802	0.0819	0.0837	0.0854
4	cos	0.9976	0.9974	0.9973	0.9972	0.9971	0.9969	0.9968	0.9966	0.9965	0.9963
	tan	0.0699	0.0717	0.0734	0.0752	0.0769	0.0787	0.0805	0.0822	0.0840	0.0857
	sin	0.0872	0.0889	0.0906	0.0924	0.0941	0.0958	0.0976	0.0993	0.1011	0.1028
5	cos	0.9962	0.9960	0.9959	0.9957	0.9956	0.9954	0.9952	0.9951	0.9949	0.9947
	tan	0.0875	0.0892	0.0910	0.0928	0.0945	0.0963	0.0981	0.0998	0.1016	0.1033
	sin	0.1045	0.1063	0.1080	0.1097	0.1115	0.1132	0.1149	0.1167	0.1184	0.1201
6	cos	0.9945	0.9943	0.9942	0.9940	0.9938	0.9936	0.9934	0.9932	0.9930	0.9928
	tan	0.1051	0.1069	0.1086	0.1104	0.1122	0.1139	0.1157	0.1175	0.1192	0.1210
	sin	0.1219	0.1236	0.1253	0.1271	0.1288	0.1305	0.1323	0.1340	0.1357	0.1374
7	cos	0.9925	0.9923	0.9921	0.9919	0.9917	0.9914	0.9912	0.9910	0.9907	0.9905
	tan	0.1228	0.1246	0.1263	0.1281	9.1299	0.1317	0.1334	0.1352	0.1370	0.1388
	sin	0.1392	0.1409	0.1426	0.1444	0.1461	0.1478	0.1495	0.1513	0.1530	0.1547
8	cos	0.9903	0.9900	0.9898	0.9895	0.9893	0.9890	0.9888	0.9885	0.9882	0.9880
	tan	0.1405	0.1423	0.1441	0.1459	0.1477	0.1495	0.1512	0.1530	0.1548	0.1566
	sin	0.1564	0.1582	0.1599	0.1616	0.1633	0.1650	0.1663	0.1685	0.1702	0.1719
9	cos	0.9877	0.9874	0.9871	0.9869	0.9866	0.9863	0.9860	0.9857	0.9854	0.9851
	tan	0.1584	0.1602	0.1620	0.1638	0.1655	0.1673	0.1691	0.1709	0.1727	0.1745
	sin	0.1736	0.1754	0.1771	0.1788	0.1805	0.1822	0.1840	0.1857	0.1874	0.1891
10	cos	0.9848	0.9845	0.9842	0.9839	0.9836	0.9833	0.9829	0.9826	0.9823	0.9820
	tan	0.1763	0.1781	0.1799	0.1817	0.1835	0.1853	0.1871	0.1890	0.1908	0.1926
	sin	0.1908	0.1925	0.1942	0.1959	0.1977	0.1994	0.2011	0.2028	0.2045	0.2062
11	cos	0.9816	0.9813	0.9810	0.9806	0.9803	0.9799	0.9796	0.9792	0.9789	0.9785
	tan	0.1944	0.1962	0.1980	0.1998	0.2016	0.2035	0.2053	0.2071	0.2089	0.2107
	sin	0.2079	0 2096	0.2113	0.2130	0.2147	0.2164	0.2181	0.2198	0.2215	0.2232
12	cos	0.9781	0.9778	0.9774	0.9770	0.9767	0.9763	0.9759	0.9755	0.9751	0.9748
	tan	0.2126	0.2144	0.2162	0.2180	0.2199	0.2217	0.2235	0.2254	0.2272	0.2290
	sin	0.2250	0.2267	0.2284	0.2300	0.2317	0.2334	0.2351	0.2368	0.2385	0.2402
13	cos	0.9744	0.9740	0.9736	0.9732	0.9728	0.9724	0.9720	0.9715	0.9711	0.9707
	tan	0.2309	0.2327	0.2345	0.2364	0.2382	0.2401	0.2419	0.2438	0.2456	0.2475
	sin	0.2419	0.2436	0.2453	0.2470	0.2487	0.2504	0.2521	0.2538	0.2554	0.2571
14	cos	0.9703	0.9699	0.9694	0.9690	0.9686	0.9681	0.9677	0.9673	0.9668	0.9664
	tan	0.2493	0.2512	0.2530	0.2549	0.2568	0.2586	0.2605	0.2623	0.2642	0.2661
	sin	0.2588	0.2605	0.2622	0.2639	0.2656	0.2672	0.2689	0.2706	0.2723	0.2740
15	cos	0.9659	0.9655	0.9650	0.9646	0.9641	0.9636	0.9632	0.9627	0.9622	0.9617
	tan	0.2679	0.2698	0.2717	0.2736	0.2754	0.2773	0.2792	0.2811	0.2830	0.2849

Trigonometric Functions

16.0°–35.9°

Angle in Degrees	Name of Function	Value of Function for Each Tenth of a Degree									
		0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
16	sin	0.2756	0.2773	0.2790	0.2807	0.2823	0 2840	0.2857	0.2874	0.2890	0.2907
	cos	0.9613	0.9608	0.9603	0.9598	0.9593	0.9588	0.9583	0.9578	0.9573	0.9568
	tan	0.2867	0.2886	0.2905	0.2924	0.2943	0.2962	0.2981	0.3000	0.3019	0.3038
17	sin	0.2924	0.2940	0.2957	0.2974	0.2990	0.3007	0.3024	0.3040	0.3057	0.3074
	cos	0.9563	0.9558	0.9553	0.9548	0 9542	0.9537	0.9532	0.9527	0.9521	0.9516
	tan	0.3057	0.3076	0.3096	0.3115	0.3134	0.3153	0.3172	0.3191	0.3211	0.3230
18	sin	0.3090	0.3107	0.3123	0.3140	0.3156	0.3173	0.3190	0.3206	0.3223	0.3239
	cos	0.9511	0.9505	0.9500	0.9494	0.9489	0.9483	0.9478	0.9472	0.9466	0.9461
	tan	0.3249	0.3269	0.3288	0.3307	0.3327	0.3346	0.3365	0.3385	0.3404	0.3424
19	sin	0.3256	0.3272	0.3289	0.3305	0.3322	0.3338	0.3355	0.3371	0.3387	0.3404
	cos	0 9455	0.9449	0.9444	0.9438	0.9432	0.9426	0.9421	0.9415	0.9409	0.9403
	tan	0.3443	0.3463	0.3482	0.3502	0.3522	0.3541	0.3561	0.3581	0.3600	0.3620
20	sin	0.3420	0.3437	0.3453	0.3469	0.3486	0.3502	0.3518	0.3535	0.3551	0.3567
	cos	0.9397	0.9391	0.9385	0.9379	0.9373	0.9367	0.9361	0.9354	0.9348	0.9342
	tan	0.3640	0.3659	0.3679	0.3699	0.3719	0.3739	0.3759	0.3779	0.3799	0.3819
21	sin	0.3584	0.3600	0.3616	0.3633	0.3649	0.3665	0.3681	0.3697	0.3714	0.3730
	cos	0.9336	0.9330	0.9323	0.9317	0.9311	0.9304	0.9298	0.9291	0.9285	0.9278
	tan	0.3839	0.3859	0.3879	0.3899	0.3919	0.3939	0.3959	0.3979	0.4000	0.4020
22	sin	0.3746	0.3762	0.3778	0.3795	0.3811	0.3827	0.3843	0.3859	0.3875	0.3891
	cos	0.9272	0.9265	0.9259	0.9252	0.9245	0.9239	0.9232	0.9225	0.9219	0.9212
	tan	0.4040	0.4061	0.4081	0.4101	0.4122	0.4142	0.4163	0.4183	0.4204	0.4224
23	sin	0.3907	0.3923	0.3939	0.3955	0.3971	0.3987	0.4003	0.4019	0.4035	0.4051
	cos	0.9205	0.9198	0.9191	0.9184	0.9178	0.9171	0.9164	0.9157	0.9150	0.9143
	tan	0.4245	0.4265	0.4286	0.4307	0.4327	0.4348	0.4369	0.4390	0.4411	0.4431
24	sin	0.4067	0.4083	0.4099	0.4115	0.4131	0.4147	0.4163	0.4179	0.4195	0.4210
	cos	0.9135	0.9128	0.9121	0.9114	0.9107	0.9100	0.9092	0.9085	0.9078	0.9070
	tan	0.4452	0.4473	0.4494	0.4515	0.4536	0.4557	0.4578	0.4599	0.4621	0.4642
25	sin	0.4226	0.4242	0.4258	0.4274	0.4289	0.4305	0.4321	0.4337	0.4352	0.4368
	cos	0.9063	0.9056	0.9048	0.9041	0.9033	0.9026	0.9018	0.9011	0.9003	0.8996
	tan	0.4663	0.4684	0.4706	0.4727	0.4748	0.4770	0.4791	0.4813	0.4834	0.4856
26	sin	0.4384	0.4399	0.4415	0.4431	0.4446	0.4462	0.4478	0.4493	0.4509	0.4524
	cos	0.8988	0.8980	0.8973	0.8965	0.8957	0.8949	0.8942	0.8934	0.8926	0.8918
	tan	0.4877	0.4899	0.4921	0.4942	0.4964	0.4986	0.5008	0.5029	0.5051	0.5073
27	sin	0.4540	0.4555	0.4571	0.4586	0.4602	0.4617	0.4633	0.4648	0.4664	0.4679
	cos	0.8910	0.8902	0.8894	0.8886	0.8878	0.8870	0.8862	0.8854	0.8846	0.8838
	tan	0.5095	0.5117	0.5139	0.5161	0.5184	0.5206	0.5228	0.5250	0.5272	0.5295
28	sin	0.4695	0.4710	0.4726	0.4741	0.4756	0.4772	0.4787	0.4802	0.4818	0.4833
	cos	0.8829	0.8821	0.8813	0.8805	0.8796	0.8788	0.8780	0.8771	0.8763	0.8755
	tan	0.5317	0.5340	0.5362	0.5384	0.5407	0.5430	0.5452	0.5475	0.5498	0.5520
29	sin	0.4848	0.4863	0.4879	0.4894	0.4909	0.4924	0.4939	0.4955	0.4970	0.4985
	cos	0.8746	0.8738	0.8729	0.8721	0.8712	0.8704	0.8695	0.8686	0.8678	0.8669
	tan	0.5543	0.5566	0.5589	0.5612	0.5635	0.5658	0.5681	0.5704	0.5727	0.5750
30	sin	0.5000	0.5015	0.5030	0.5045	0.5060	0.5075	0.5090	0.5105	0.5120	0.5135
	cos	0.8660	0.8652	0.8643	0.8634	0.8625	0.8616	0 8607	0.8599	0.8590	0.8581
	tan	0.5774	0.5797	0.5820	0.5844	0.5867	0.5890	0.5914	0.5938	0.5961	0.5985
31	sin	0.5150	0.5165	0.5180	0.5195	0.5210	0.5225	0.5240	0.5255	0.5270	0.5284
	cos	0.8572	0.8563	0.8554	0.8545	0.8536	0.8526	0.8517	0.8508	0.8499	0.8490
	tan	0.6009	0.6032	0.6056	0.6080	0.6104	0.6128	0.6152	0.6176	0.6200	0.6224
32	sin	0.5299	0.5314	0.5329	0.5344	0.5358	0.5373	0.5388	0.5402	0.5417	0.5432
	cos	0.8480	0.8471	0.8462	0.8453	0.8443	0.8434	0.8425	0.8415	0.8406	0.8396
	tan	0.6249	0.6273	0.6297	0.6322	0.6346	0.6371	0.6395	0.6420	0.6445	0.6469
33	sin	0.5446	0.5461	0.5476	0.5490	0.5505	0.5519	0.5534	0.5548	0.5563	0.5577
	cos	0.8387	0.8377	0.8368	0.8358	0.8348	0.8339	0.8329	0.8320	0.8310	0.8300
	tan	0.6494	0.6519	0.6544	0.6569	0.6594	0.6619	0.6644	0.6669	0.6694	0.6720
34	sin	0.5592	0.5606	0.5621	0.5635	0.5650	0.5664	0.5678	0.5693	0.5707	0.5721
	cos	0.8290	0.8281	0.8271	0.8261	0.8251	0.8241	0.8231	0.8221	0.8211	0.8202
	tan	0.6745	0.6771	0.6796	0.6822	0.6847	0.6873	0.6899	0.6924	0.6950	0.6976
35	sin	0.5736	0.5750	0.5764	0.5779	0.5793	0.5807	0.5821	0.5835	0.5850	0.5864
	cos	0.8192	0.8181	0.8171	0.8161	0.8151	0.8141	0.8131	0.8121	0.8111	0.8100
	tan	0.7002	0.7028	0.7054	0.7080	0.7107	0.7133	0.7159	0.7186	0.7212	0.7239

Trigonometric Functions

36.0°–55.9°

Angle in Degrees	Name of Function	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	sin	0.5878	0.5892	0.5906	0.5920	0.5934	0.5948	0.5962	0.5976	0.5990	0.6004
36	cos	0.8090	0.8080	0.8070	0.8059	0.8049	0.8039	0.8028	0.8018	0.8007	0.7997
	tan	0.7265	0.7292	0.7319	0.7346	0.7373	0.7400	0.7427	0.7454	0.7481	0.7508
	sin	0.6018	0.6032	0.6046	0.6060	0.6074	0.6088	0.6101	0.6115	0.6129	0.6143
37	cos	0.7986	0.7976	0 7965	0.7955	0.7944	0.7934	0.7923	0.7912	0.7902	0.7891
	tan	0.7536	0.7563	0.7590	0.7618	0.7646	0.7673	0.7701	0.7729	0.7757	0.7785
	sin	0.6157	0.6170	0.6184	0.6198	0.6211	0.6225	0.6239	0.6252	0.6266	0.6280
38	cos	0.7880	0.7869	0.7859	0.7848	0.7837	0.7826	0.7815	0.7804	0.7793	0.7782
	tan	0.7813	0.7841	0.7869	0.7898	0.7926	0.7954	0.7983	0.8012	0.8040	0.8069
	sin	0.6293	0.6307	0.6320	0.6334	0.6347	0.6361	0.6374	0.6388	0.6401	0.6414
39	cos	0.7771	0.7760	0.7749	0.7738	0.7727	0.7716	0.7705	0.7694	0.7683	0.7672
	tan	0.8098	0.8127	0.8156	0.8185	0.8214	0.8243	0.8273	0.8302	0.8332	0.8361
	sin	0.6428	0.6441	0.6455	0.6468	0.6481	0.6494	0.6508	0.6521	0.6534	0.6547
40	cos	0.7660	0.7649	0.7638	0.7627	0.7615	0.7604	0.7593	0.7581	0.7570	0.7559
	tan	0.8391	0.8421	0.8451	0.8481	0.8511	0.8541	0.8571	0.8601	0.8632	0.8662
	sin	0.6561	0.6574	0.6587	0.6600	0.6613	0.6626	0.6639	0.6653	0.6665	0.6678
41	cos	0.7547	0.7536	0.7524	0.7513	0.7501	0.7490	0.7478	0.7466	0.7455	0.7443
	tan	0.8693	0.8724	0.8754	0.8785	0.8816	0.8847	0.8878	0.8910	0.8941	0.8972
	sin	0.6691	0.6704	0.6717	0.6730	0.6743	0.6756	0.6769	0.6782	0.6794	0.6807
42	cos	0.7431	0.7420	0.7408	0.7396	0.7385	0.7373	0.7361	0.7349	0.7337	0.7325
	tan	0.9004	0.9036	0.9067	0.9099	0.9131	0.9163	0.9195	0.9228	0.9260	0.9293
	sin	0.6820	0.6833	0.6845	0.6858	0.6871	0.6884	0.6896	0.6909	0.6921	0.6934
43	cos	0.7314	0.7302	0.7290	0.7278	0.7266	0.7254	0.7242	0.7230	0.7218	0.7206
	tan	0.9325	0.9358	0.9391	0.9424	0.9457	0.9490	0.9523	0.9556	0.9590	0.9623
	sin	0.6947	0.6959	0.6972	0.6984	0.6997	0.7009	0.7022	0.7034	0.7046	0.7059
44	cos	0.7193	0.7181	0.7169	0.7157	0.7145	0.7133	0.7120	0.7108	0.7096	0.7083
	tan	0.9657	0.9691	0.9725	0.9759	0.9793	0.9827	0.9861	0.9896	0.9930	0.9965
	sin	0.7071	0.7083	0.7096	0.7108	0.7120	0.7133	0.7145	0.7157	0.7169	0.7181
45	cos	0.7071	0.7059	0.7046	0.7034	0.7022	0.7009	0.6997	0.6984	0.6972	0.6959
	tan	1.0000	1.0035	1.0070	1.0105	1.0141	1.0176	1.0212	1.0247	1.0283	1.0319
	sin	0.7193	0.7206	0.7218	0.7230	0.7242	0.7254	0.7266	0.7278	0.7290	0.7302
46	cos	0.6947	0.6934	0.6921	0.6909	0.6896	0.6884	0.6871	0.6858	0.6845	0.6833
	tan	1.0355	1.0392	1.0428	1.0464	1.0501	1.0538	1.0575	1.0612	1.0649	1.0686
	sin	0.7314	0.7325	0.7337	0.7349	0.7361	0.7373	0.7385	0.7396	0.7408	0.7420
47	cos	0.6820	0.6807	0.6794	0.6782	0.6769	0.6756	0.6743	0.6730	0.6717	0.6704
	tan	1.0724	1.0761	1.0799	1.0837	1.0875	1.0913	1.0951	1.0990	1.1028	1.1067
	sin	0.7431	0.7443	0.7455	0.7466	0.7478	0.7490	0.7501	0.7513	0.7524	0.7536
48	cos	0.6691	0.6678	0.6665	0.6652	0.6639	0.6626	0.6613	0.6600	0.6587	0.6574
	tan	1.1106	1.1145	1.1184	1.1224	1.1263	1.1303	1.1343	1.1383	1.1423	1.1463
	sin	0.7547	0.7559	0.7570	0.7581	0.7593	0.7604	0.7615	0.7627	0.7638	0.7649
49	cos	0.6561	0.6547	0.6534	0.6521	0.6508	0.6494	0.6481	0.6468	0.6455	0.6441
	tan	1.1504	1.1544	1.1585	1.1626	1.1667	1.1708	1.1750	1.1792	1.1833	1.1875
	sin	0.7660	0.7672	0.7683	0.7694	0.7705	0.7716	0.7727	0.7738	0.7749	0.7760
50	cos	0.6428	0.6414	0.6401	0.6388	0.6374	0.6361	0.6347	0.6334	0.6320	0.6307
	tan	1.1918	1.1960	1.2002	1.2045	1.2088	1.2131	1.2174	1.2218	1.2261	1.2305
	sin	0.7771	0.7782	0.7793	0.7804	0.7815	0.7826	0.7837	0.7848	0.7859	0.7869
51	cos	0.6293	0.6280	0.6266	0.6252	0.6239	0.6225	0.6211	0.6198	0.6184	0.6170
	tan	1.2349	1.2393	1.2437	1.2482	1.2527	1.2572	1.2617	1.2662	1.2708	1.2753
	sin	0.7880	0.7891	0.7902	0.7912	0.7923	0.7934	0.7944	0.7955	0.7965	0.7976
52	cos	0.6157	0.6143	0.6129	0.6115	0.6101	0.6088	0.6074	0.6060	0.6046	0.6032
	tan	1.2799	1.2846	1.2892	1.2938	1.2985	1.3032	1.3079	1.3127	1.3175	1.3222
	sin	0.7986	0.7997	0.8007	0.8018	0.8028	0.8039	0.8049	0.8059	0.8070	0.8080
53	cos	0.6018	0.6004	0.5990	0.5976	0.5962	0.5948	0.5934	0.5920	0.5906	0.5892
	tan	1.3270	1.3319	1.3367	1.3416	1.3465	1.3514	1.3564	1.3613	1.3663	1.3713
	sin	0.8090	0.8100	0.8111	0.8121	0.8131	0.8141	0.8151	0.8161	0.8171	0.8181
54	cos	0.5878	0.5864	0.5850	0.5835	0.5821	0.5807	0.5793	0.5779	0.5764	0.5750
	tan	1.3764	1.3814	1.3865	1.3916	1.3968	1.4019	1.4071	1.4124	1.4176	1.4229
	sin	0.8192	0.8202	0.8211	0.8221	0.8231	0.8241	0.8251	0.8261	0.8271	0.8281
55	cos	0.5736	0.5721	0.5707	0.5693	0.5678	0.5664	0.5650	0.5635	0.5621	0.5606
	tan	1.4281	1.4335	1.4388	1.4442	1.4496	1.4550	1.4605	1.4659	1.4715	1.4770

Trigonometric Functions

56.0°–75.9°

Angle in Degrees	Name of Function	Value of Function for Each Tenth of a Degree									
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
56	sin	0.8290	0.8300	0.8310	0.8320	0.8329	0.8339	0.8348	0.8358	0.8368	0.8377
	cos	0.5592	0.5577	0.5563	0.5548	0.5534	0.5519	0.5505	0.5490	0.5476	0.5461
	tan	1.4826	1.4882	1.4938	1.4994	1.5051	1.5108	1.5166	1.5224	1.5282	1.5340
57	sin	0.8387	0.8396	0.8406	0.8415	0.8425	0.8434	0.8443	0.8453	0.8462	0.8471
	cos	0.5446	0.5432	0.5417	0.5402	0.5388	0.5373	0.5358	0.5344	0.5329	0.5314
	tan	1.5399	1.5458	1.5517	1.5577	1.5637	1.5697	1.5757	1.5818	1.5880	1.5941
58	sin	0.8480	0.8490	0.8499	0.8508	0.8517	0.8526	0.8536	0.8545	0.8554	0.8563
	cos	0.5299	0.5284	0.5270	0.5255	0.5240	0.5225	0.5210	0.5195	0.5180	0.5165
	tan	1.6003	1.6066	1.6128	1.6191	1.6255	1.6319	1.6383	1.6447	1.6512	1.6577
59	sin	0.8572	0.8581	0.8590	0.8599	0.8607	0.8616	0.8625	0.8634	0.8643	0.8652
	cos	0.5150	0.5135	0.5120	0.5105	0.5090	0.5075	0.5060	0.5045	0.5030	0.5015
	tan	1.6643	1.6709	1.6775	1.6842	1.6909	1.6977	1.7045	1.7113	1.7182	1.7251
60	sin	0.8660	0.8669	0.8678	0.8686	0.8695	0.8704	0.8712	0.8721	0.8729	0.8738
	cos	0.5000	0.4985	0.4970	0.4955	0.4939	0.4924	0.4909	0.4894	0.4879	0.4863
	tan	1.7321	1.7391	1.7461	1.7532	1.7603	1.7675	1.7747	1.7820	1.7893	1.7966
61	sin	0.8746	0.8755	0.8763	0.8771	0.8780	0.8788	0.8796	0.8805	0.8813	0.8821
	cos	0.4848	0.4833	0.4818	0.4802	0.4787	0.4772	0.4756	0.4741	0.4726	0.4710
	tan	1.8040	1.8115	1.8190	1.8265	1.8341	1.8418	1.8495	1.8572	1.8650	1.8728
62	sin	0.8829	0.8838	0.8846	0.8854	0.8862	0.8870	0.8878	0.8886	0.8894	0.8902
	cos	0.4695	0.4679	0.4664	0.4648	0.4633	0.4617	0.4602	0.4586	0.4571	0.4555
	tan	1.8807	1.8887	1.8967	1.9047	1.9128	1.9210	1.9292	1.9375	1.9458	1.9542
63	sin	0.8910	0.8918	0.8926	0.8934	0.8942	0.8949	0.8957	0.8965	0.8973	0.8980
	cos	0.4540	0.4524	0.4509	0.4493	0.4478	0.4462	0.4446	0.4431	0.4415	0.4399
	tan	1.9626	1.9711	1.9797	1.9883	1.9970	2.0057	2.0145	2.0233	2.0323	2.0413
64	sin	0.8988	0.8996	0.9003	0.9011	0.9018	0.9026	0.9033	0.9041	0.9048	0.9056
	cos	0.4384	0.4368	0.4352	0.4337	0.4321	0.4305	0.4289	0.4274	0.4258	0.4242
	tan	2.0503	2.0594	2.0686	2.0778	2.0872	2.0965	2.1060	2.1155	2.1251	2.1348
65	sin	0.9063	0.9070	0.9078	0.9085	0.9092	0.9100	0.9107	0.9114	0.9121	0.9128
	cos	0.4226	0.4210	0.4195	0.4179	0.4163	0.4147	0.4131	0.4115	0.4099	0.4083
	tan	2.1445	2.1543	2.1642	2.1742	2.1842	2.1943	2.2045	2.2148	2.2251	2.2355
66	sin	0.9135	0.9143	0.9150	0.9157	0.9164	0.9171	0.9178	0.9184	0.9191	0.9198
	cos	0.4067	0.4051	0.4035	0.4019	0.4003	0.3987	0.3971	0.3955	0.3939	0.3923
	tan	2.2460	2.2566	2.2673	2.2781	2.2889	2.2998	2.3109	2.3220	2.3332	2.3445
67	sin	0.9205	0.9212	0.9219	0.9225	0.9232	0.9239	0.9245	0.9252	0.9259	0.9265
	cos	0.3907	0.3891	0.3875	0.3859	0.3843	0.3827	0.3811	0.3795	0.3778	0.3762
	tan	2.3559	2.3673	2.3789	2.3906	2.4023	2.4142	2.4262	2.4383	2.4504	2.4627
68	sin	0.9272	0.9278	0.9285	0.9291	0.9298	0.9304	0.9311	0.9317	0.9323	0.9330
	cos	0.3746	0.3730	0.3714	0.3697	0.3681	0.3665	0.3649	0.3633	0.3616	0.3600
	tan	2.4751	2.4876	2.5002	2.5129	2.5257	2.5386	2.5517	2.5649	2.5782	2.5916
69	sin	0.9336	0.9342	0.9348	0.9354	0.9361	0.9367	0.9373	0.9379	0.9385	0.9391
	cos	0.3584	0.3567	0.3551	0.3535	0.3518	0.3502	0.3486	0.3469	0.3453	0.3437
	tan	2.6051	2.6187	2.6325	2.6464	2.6605	2.6746	2.6889	2.7034	2.7179	2.7326
70	sin	0.9397	0.9403	0.9409	0.9415	0.9421	0.9426	0.9432	0.9438	0.9444	0.9449
	cos	0.3420	0.3404	0.3387	0.3371	0.3355	0.3338	0.3322	0.3305	0.3289	0.3272
	tan	2.7475	2.7625	2.7776	2.7929	2.8083	2.8239	2.8397	2.8556	2.8716	2.8878
71	sin	0.9455	0.9461	0.9466	0.9472	0.9478	0.9483	0.9489	0.9494	0.9500	0.9505
	cos	0.3256	0.3239	0.3223	0.3206	0.3190	0.3173	0.3156	0.3140	0.3123	0.3107
	tan	2.9042	2.9208	2.9375	2.9544	2.9714	2.9887	3.0061	3.0237	3.0415	3.0595
72	sin	0.9511	0.9516	0.9521	0.9527	0.9532	0.9537	0.9542	0.9548	0.9553	0.9558
	cos	0.3090	0.3074	0.3057	0.3040	0.3024	0.3007	0.2990	0.2974	0.2957	0.2940
	tan	3.0777	3.0961	3.1146	3.1334	3.1524	3.1716	3.1910	3.2106	3.2305	3.2506
73	sin	0.9563	0.9568	0.9573	0.9578	0.9583	0.9588	0.9593	0.9598	0.9603	0.9608
	cos	0.2924	0.2907	0.2890	0.2874	0.2857	0.2840	0.2823	0.2807	0.2790	0.2773
	tan	3.2709	3.2914	3.3122	3.3332	3.3544	3.3759	3.3977	3.4197	3.4420	3.4646
74	sin	0.9613	0.9617	0.9622	0.9627	0.9632	0.9636	0.9641	0.9646	0.9650	0.9655
	cos	0.2756	0.2740	0.2723	0.2706	0.2689	0.2672	0.2656	0.2639	0.2622	0.2605
	tan	3.4874	3.5105	3.5339	3.5576	3.5816	3.6059	3.6305	3.6554	3.6806	3.7062
75	sin	0.9659	0.9664	0.9668	0.9673	0.9677	0.9681	0.9686	0.9690	0.9694	0.9699
	cos	0.2588	0.2571	0.2554	0.2538	0.2521	0.2504	0.2487	0.2470	0.2453	0.2436
	tan	3.7321	3.7583	3.7848	3.8118	3.8391	3.8667	3.8947	3.9232	3.9520	3.9812

Trigonometric Functions **76.0°–89.9°**

Angle in Degrees	Name of Function	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
		Value of Function for Each Tenth of a Degree									
	sin	0.9703	0.9707	0.9711	0.9715	0.9720	0.9724	0.9728	0.9732	0.9736	0.9740
76	cos	0.2419	0.2402	0.2385	0.2368	0.2351	0.2334	0.2317	0.2300	0.2284	0.2267
	tan	4.0108	4.0408	4.0713	4.1022	4.1335	4.1653	4.1976	4.2303	4.2635	4.2972
	sin	0.9744	0.9748	0.9751	0.9755	0.9759	0.9763	0.9767	0.9770	0.9774	0.9778
77	cos	0.2250	0.2232	0.2215	0.2198	0.2181	0.2164	0.2147	0.2130	0.2113	0.2096
	tan	4.3315	4.3662	4.4015	4.4374	4.4737	4.5107	4.5483	4.5864	4.6252	4.6646
	sin	0.9781	0.9785	0.9789	0.9792	0.9796	0.9799	0.9803	0.9806	0.9810	0.9813
78	cos	0.2079	0.2062	0.2045	0.2028	0.2011	0.1994	0.1977	0.1959	0.1942	0.1925
	tan	4.7046	4.7453	4.7867	4.8288	4.8716	4.9152	4.9594	5.0045	5.0504	5.0970
	sin	0.9816	0.9820	0.9823	0.9826	0.9829	0.9833	0.9836	0.9839	0.9842	0.9845
79	cos	0.1908	0.1891	0.1874	0.1857	0.1840	0.1822	0.1805	0.1788	0.1771	0.1754
	tan	5.1446	5.1929	5.2422	5.2924	5.3435	5.3955	5.4486	5.5026	5.5578	5.6140
	sin	0.9848	0.9851	0.9854	0.9857	0.9860	0.9863	0.9866	0.9869	0.9871	0.9874
80	cos	0.1736	0.1719	0.1702	0.1685	0.1668	0.1650	0.1633	0.1616	0.1599	0.1582
	tan	5.6713	5.7297	5.7894	5.8502	5.9124	5.9758	6.0405	6.1066	6.1742	6.2432
	sin	0.9877	0.9880	0.9882	0.9885	0.9888	0.9890	0.9893	0.9895	0.9898	0.9900
81	cos	0.1564	0.1547	0.1530	0.1513	0.1495	0.1478	0.1461	0.1444	0.1426	0.1409
	tan	6.3138	6.3859	6.4596	6.5350	6.6122	6.6912	6.7720	6.8548	6.9395	7.0264
	sin	0.9903	0.9905	0.9907	0.9910	0.9912	0.9914	0.9917	0.9919	0.9921	0.9923
82	cos	0.1392	0.1374	0.1357	0.1340	0.1323	0.1305	0.1288	0.1271	0.1253	0.1236
	tan	7.1154	7.2066	7.3002	7.3962	7.4947	7.5958	7.6996	7.8062	7.9158	8.0285
	sin	0.9925	0.9928	0.9930	0.9932	0.9934	0.9936	0.9938	0.9940	0.9942	0.9943
83	cos	0.1219	0.1201	0.1184	0.1167	0.1149	0.1132	0.1115	0.1097	0.1080	0.1063
	tan	8.1443	8.2636	8.3863	8.5126	8.6427	8.7769	8.9152	9.0579	9.2052	9.3572
	sin	0.9945	0.9947	0.9949	0.9951	0.9952	0.9954	0.9956	0.9957	0.9959	0.9960
84	cos	0.1045	0.1028	0.1011	0.0993	0.0976	0.0958	0.0941	0.0924	0.0906	0.0889
	tan	9.5144	9.6768	9.8448	10.02	10.20	10.39	10.58	10.78	10.99	11.20
	sin	0.9962	0.9963	0.9965	0.9966	0.9968	0.9969	0.9971	0.9972	0.9973	0.9974
85	cos	0.0872	0.0854	0.0837	0.0819	0.0802	0.0785	0.0767	0.0750	0.0732	0.0715
	tan	11.43	11.66	11.91	12.16	12.43	12.71	13.00	13.30	13.62	13.95
	sin	0.9976	0.9977	0.9978	0.9979	0.9980	0.9981	0.9982	0.9983	0.9984	0.9985
86	cos	0.0698	0.0680	0.0663	0.0645	0.0628	0.0610	0.0593	0.0576	0.0558	0.0541
	tan	14.30	14.67	15.06	15.89	15.46	16.35	16.83	17.34	17.89	18.46
	sin	0.9986	0.9987	0.9988	0.9989	0.9990	0.9990	0.9991	0.9992	0.9993	0.9993
87	cos	0.0523	0.0506	0.0488	0.0471	0.0454	0.0436	0.0419	0.0401	0.0384	0.0366
	tan	19.08	19.74	20.45	21.20	22.02	22.90	23.86	24.90	26.03	27.27
	sin	0.9994	0.9995	0.9995	0.9996	0.9996	0.9997	0.9997	0.9997	0.9998	0.9998
88	cos	0.0349	0.0332	0.0314	0.0297	0.0279	0.0262	0.0244	0.0227	0.0209	0.0192
	tan	28.64	30.14	31.82	33.69	35.80	38.19	40.92	44.07	47.74	52.08
	sin	0.9998	0.9999	0.9999	0.9999	0.9999	1.000	1.000	1.000	1.000	1.000
89	cos	0.0175	0.0157	0.0140	0.0122	0.0105	0.0087	0.0070	0.0052	0.0035	0.0017
	tan	57.29	63.66	71.62	81.85	95.49	114.6	143.2	191.0	286.5	573.0

13. EXPONENTIAL AND HYPERBOLIC TABLES

The following tables give values of e^x, e^{-x}, $\sinh x$, $\cosh x$ and $\tanh x$ for values of x from 0.00 to 6.00 in intervals of 0.01.

To facilitate computations involving multiplication, the common logarithms of e^x, $\sinh x$, $\cosh x$, and $\tanh x$ are also given.

For values of x greater than 6, e^x may be computed from the relationship $e^x = \log^{-1} (x \log_{10} e) = \log^{-1} 0.43429x$; e^{-x} approaches zero; $\sinh x$ and $\cosh x$ are approximately equal and become $0.5\,e^x$; and $\tanh x$ and $\coth x$ have values approximately equal to unity.

Where more accurate values of the exponentials and functions are required they may be computed from the following relationships.

$$e = 2.71828\ 18285 \qquad \frac{1}{e} = 0.36787\ 94412$$

$$M = \log_{10} e = 0.43429\ 44819 \qquad \frac{1}{M} = \log_e 10 = 2.30258\ 50930$$

$$e^x = \log^{-1} Mx \qquad e^{-x} = \log^{-1} - Mx$$

$$\sinh x = \frac{e^x - e^{-x}}{2} \qquad \cosh x = \frac{e^x + e^{-x}}{2} \qquad \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\operatorname{csch} x = \frac{1}{\sinh x} \qquad \operatorname{sech} x = \frac{1}{\cosh x} \qquad \coth x = \frac{1}{\tanh x}$$

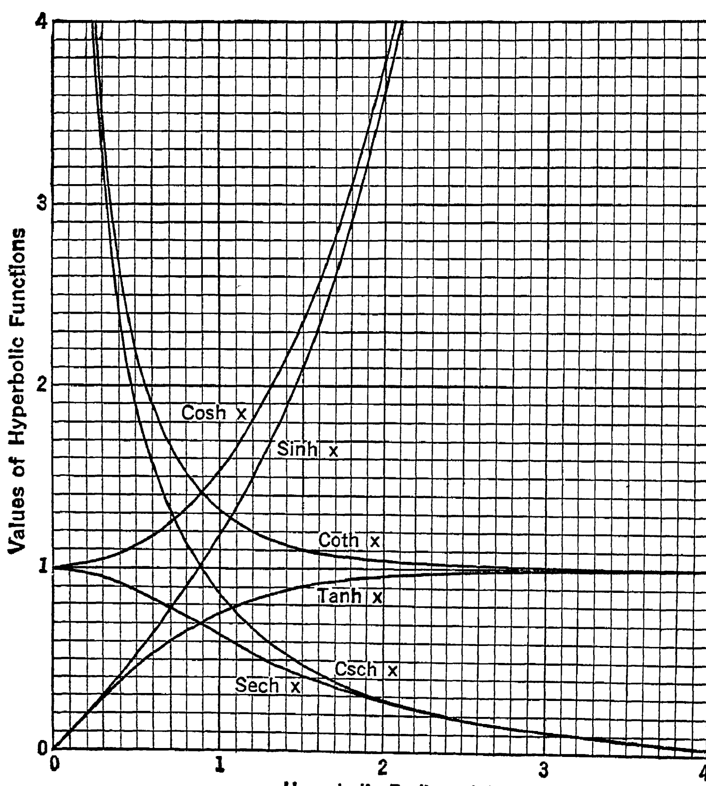


Chart of the Hyperbolic Functions.

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
0.00	**1.0000**	**1.0000**	**0.0000**	**1.0000**	**.00000**	**0.00000**	**$-\infty$**	**0.00000**	**$-\infty$**
0.01	1.0101	.99005	0.0100	1.0001	.01000	.00434	$\bar{2}$.00001	.00002	$\bar{3}$.99999
0.02	1.0202	.98020	0.0200	1.0002	.02000	.00869	.30106	.00009	$\bar{2}$.30097
0.03	1.0305	.97045	0.0300	1.0005	.02999	.01303	.47719	.00020	.47699
0.04	1.0408	.96079	0.0400	1.0008	.03998	.01737	.60218	.00035	.60183
0.05	1.0513	.95123	0.0500	1.0013	.04996	.02171	.69915	.00054	.69861
0.06	1.0618	.94176	0.0600	1.0018	.05993	.02606	.77841	.00078	.77763
0.07	1.0725	.93239	0.0701	1.0025	.06989	.03040	.84545	.00106	.84439
0.08	1.0833	.92312	0.0801	1.0032	.07983	.03474	.90355	.00139	.90216
0.09	1.0942	.91393	0.0901	1.0041	.08976	.03909	.95483	.00176	.95307
0.10	**1.1052**	**.90484**	**0.1002**	**1.0050**	**.09967**	**0.04343**	**$\bar{1}$.00072**	**0.00217**	**$\bar{2}$.99856**
0.11	1.1163	.89583	0.1102	1.0061	.10956	.04777	.04227	.00262	$\bar{1}$.03965
0.12	1.1275	.88692	0.1203	1.0072	.11943	.05212	.08022	.00312	.07710
0.13	1.1388	.87810	0.1304	1.0085	.12927	.05646	.11517	.00366	.11151
0.14	1.1503	.86936	0.1405	1.0098	.13909	.06080	.14755	.00424	.14330
0.15	1.1618	.86071	0.1506	1.0113	.14889	.06514	.17772	.00487	.17285
0.16	1.1735	.85214	0.1607	1.0128	.15865	.06949	.20597	.00554	.20044
0.17	1.1853	.84366	0.1708	1.0145	.16838	.07383	.23254	.00625	.22629
0.18	1.1972	.83527	0.1810	1.0162	.17808	.07817	.25762	.00700	.25062
0.19	1.2092	.82696	0.1911	1.0181	.18775	.08252	.28136	.00779	.27357
0.20	**1.2214**	**.81873**	**0.2013**	**1.0201**	**.19738**	**0.08686**	**$\bar{1}$.30392**	**0.00863**	**1.29529**
0.21	1.2337	.81058	0.2115	1.0221	.20697	.09120	.32541	.00951	.31590
0.22	1.2461	.80252	0.2218	1.0243	.21652	.09554	.34592	.01043	.33549
0.23	1.2586	.79453	0.2320	1.0266	.22603	.09989	.36555	.01139	.35416
0.24	1.2712	.78663	0.2423	1.0289	.23550	.10423	.38437	.01239	.37198
0.25	1.2840	.77880	0.2526	1.0314	.24492	.10857	.40245	.01343	.38902
0.26	1.2969	.77105	0.2629	1.0340	.25430	.11292	.41986	.01452	.40534
0.27	1.3100	.76338	0.2733	1.0367	.26362	.11726	.43663	.01564	.42099
0.28	1.3231	.75578	0.2837	1.0395	.27291	.12160	.45282	.01681	.43601
0.29	1.3364	.74826	0.2941	1.0423	.28213	.12595	.46847	.01801	.45046
0.30	**1.3499**	**.74082**	**0.3045**	**1.0453**	**.29131**	**0.13029**	**1.48362**	**0.01926**	**$\bar{1}$.46436**
0.31	1.3634	.73345	0.3150	1.0484	.30044	.13463	.49830	.02054	.47775
0.32	1.3771	.72615	0.3255	1.0516	.30951	.13897	.51254	.02107	.49067
0.33	1.3910	.71892	0.3360	1.0549	.31852	.14332	.52637	.02323	.50314
0.34	1.4049	.71177	0.3466	1.0584	.32748	.14766	.53981	.02463	.51518
0.35	1.4191	.70469	0.3572	1.0619	.33638	.15200	.55290	.02607	.52682
0.36	1.4333	.69768	0.3678	1.0655	.34521	.15635	.56564	.02755	.53809
0.37	1.4477	.69073	0.3785	1.0692	.35399	.16069	.57807	.02907	.54899
0.38	1.4623	.68386	0.3892	1.0731	.36271	.16503	.59019	.03063	.55956
0.39	1.4770	.67706	0.4000	1.0770	.37136	.16937	.60202	.03222	.56980
0.40	**1.4918**	**.67032**	**0.4108**	**1.0811**	**.37995**	**0.17372**	**$\bar{1}$.61358**	**0.03385**	**$\bar{1}$.57973**
0.41	1.5068	.66365	0.4216	1.0852	.38847	.17806	.62488	.03552	.58936
0.42	1.5220	.65705	0.4325	1.0895	.39693	.18240	.63594	.03723	.59871
0.43	1.5373	.65051	0.4434	1.0939	.40532	.18675	.64677	.03897	.60780
0.44	1.5527	.64404	0.4543	1.0984	.41364	.19109	.65738	.04075	.61663
0.45	1.5683	.63763	0.4653	1.1030	.42190	.19543	.66777	.04256	.62521
0.46	1.5841	.63128	0.4764	1.1077	.43008	.19978	.67797	.04441	.63355
0.47	1.6000	.62500	0.4875	1.1125	.43820	.20412	.68797	.04630	.64167
0.48	1.6161	.61878	0.4986	1.1174	.44624	.20846	.69779	.04822	.64957
0.49	1.6323	.61263	0.5098	1.1225	.45422	.21280	.70744	.05018	.65726
0.50	**1.6487**	**.60653**	**0.5211**	**1.1276**	**.46212**	**0.21715**	**1.71692**	**0.05217**	**1.66475**
0.51	1.6653	.60050	0.5324	1.1329	.46995	.22149	.72624	.05419	.67205
0.52	1.6820	.59452	0.5438	1.1383	.47770	.22583	.73540	.05625	.67916
0.53	1.6989	.58860	0.5552	1.1438	.48538	.23018	.74442	.05834	.68608
0.54	1.7160	.58275	0.5666	1.1494	.49299	.23452	.75330	.06046	.69284
0.55	1.7333	.57695	0.5782	1.1551	.50052	.23886	.76204	.06262	.69942
0.56	1.7507	.57121	0.5897	1.1609	.50798	.24320	.77065	.06481	.70584
0.57	1.7683	.56553	0.6014	1.1669	.51536	.24755	.77914	.06703	.71211
0.58	1.7860	.55990	0.6131	1.1730	.52267	.25189	.78751	.06929	.71822
0.59	1.8040	.55433	0.6248	1.1792	.52990	.25623	.79576	.07157	.72419
0.60	**1.8221**	**.54881**	**0.6367**	**1.1855**	**.53705**	**0.26058**	**1.80390**	**0.07389**	**$\bar{1}$.73001**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
0.60	**1.8221**	**.54881**	**0.6367**	**1.1855**	**.53705**	**0.26058**	**$\bar{1}$.80390**	**0.07389**	**$\bar{1}$.73001**
0.61	1.8404	.54335	0.6485	1.1919	.54413	.26492	.81194	.07624	.73570
0.62	1.8589	.53794	0.6605	1.1984	.55113	.26926	.81987	.07861	.74125
0.63	1.8776	.53259	0.6725	1.2051	.55805	.27361	.82770	.08102	.74667
0.64	1.8965	.52729	0.6846	1.2119	.56490	.27795	.83543	.08346	.75197
0.65	1.9155	.52205	0.6967	1.2188	.57167	.28229	.84308	.08593	.75715
0.66	1.9348	.51685	0.7090	1.2258	.57836	.28663	.85063	.08843	.76220
0.67	1.9542	.51171	0.7213	1.2330	.58498	.29098	.85809	.09095	.76714
0.68	1.9739	.50662	0.7336	1.2402	.59152	.29532	.86548	.09351	.77197
0.69	1.9937	.50158	0.7461	1.2476	.59798	.29966	.87278	.09609	.77669
0.70	**2.0138**	**.49659**	**0.7586**	**1.2552**	**.60437**	**0.30401**	**$\bar{1}$.88000**	**0.09870**	**$\bar{1}$.78130**
0.71	2.0340	.49164	0.7712	1.2628	.61068	.30835	.88715	.10134	.78581
0.72	2.0544	.48675	0.7838	1.2706	.61691	.31269	.89423	.10401	.79022
0.73	2.0751	.48191	0.7966	1.2785	.62307	.31704	.90123	.10670	.79453
0.74	2.0959	.47711	0.8094	1.2865	.62915	.32138	.90817	.10942	.79875
0.75	2.1170	.47237	0.8223	1.2947	.63515	.32572	.91504	.11216	.80288
0.76	2.1383	.46767	0.8353	1.3030	.64108	.33006	.92185	.11493	.80691
0.77	2.1598	.46301	0.8484	1.3114	.64693	.33441	.92859	.11773	.81086
0.78	2.1815	.45841	0.8615	1.3199	.65271	.33875	.93527	.12055	.81472
0.79	2.2034	.45384	0.8748	1.3286	.65841	.34309	.94190	.12340	.81850
0.80	**2.2255**	**.44933**	**0.8881**	**1.3374**	**.66404**	**0.34744**	**$\bar{1}$.94846**	**0.12627**	**1.82219**
0.81	2.2479	.44486	0.9015	1.3464	.66959	.35178	.95498	.12917	.82581
0.82	2.2705	.44043	0.9150	1.3555	.67507	.35612	.96144	.13209	.82935
0.83	2.2933	.43605	0.9286	1.3647	.68048	.36046	.96784	.13503	.83281
0.84	2.3164	.43171	0.9423	1.3740	.68581	.36481	.97420	.13800	.83620
0.85	2.3396	.42741	0.9561	1.3835	.69107	.36915	.98051	.14099	.83952
0.86	2.3632	.42316	0.9700	1.3932	.69626	.37349	.98677	.14400	.84277
0.87	2.3869	.41895	0.9840	1.4029	.70137	.37784	.99299	.14704	.84595
0.88	2.4109	.41478	0.9981	1.4128	.70642	.38218	.99916	.15009	.84906
0.89	2.4351	.41066	1.0122	1.4229	.71139	.38652	0.00528	.15317	.85211
0.90	**2.4596**	**.40657**	**1.0265**	**1.4331**	**.71630**	**0.39087**	**0.01137**	**0.15627**	**$\bar{1}$.85509**
0.91	2.4843	.40252	1.0409	1.4434	.72113	.39521	.01741	.15939	.85801
0.92	2.5093	.39852	1.0554	1.4539	.72590	.39955	.02341	.16254	.86088
0.93	2.5345	.39455	1.0700	1.4645	.73059	.40389	.02937	.16570	.86368
0.94	2.5600	.39063	1.0847	1.4753	.73522	.40824	.03530	.16888	.86642
0.95	2.5857	.38674	1.0995	1.4862	.73978	.41258	.04119	.17208	.86910
0.96	2.6117	.38289	1.1144	1.4973	.74428	.41692	.04704	.17531	.87173
0.97	2.6379	.37908	1.1294	1.5085	.74870	.42127	.05286	.17855	.87431
0.98	2.6645	.37531	1.1446	1.5199	.75307	.42561	.05864	.18181	.87683
0.99	2.6912	.37158	1.1598	1.5314	.75736	.42995	.06439	.18509	.87930
1.00	**2.7183**	**.36788**	**1.1752**	**1.5431**	**.76159**	**0.43429**	**0.07011**	**0.18839**	**1.88172**
1.01	2.7456	.36422	1.1907	1.5549	.76576	.43864	.07580	.19171	.88409
1.02	2.7732	.36059	1.2063	1.5669	.76987	.44298	.08146	.19504	.88642
1.03	2.8011	.35701	1.2220	1.5790	.77391	.44732	.08708	.19839	.88869
1.04	2.8292	.35345	1.2379	1.5913	.77789	.45167	.09268	.20176	.89092
1.05	2.8577	.34994	1.2539	1.6038	.78181	.45601	.09825	.20515	.89310
1.06	2.8864	.34646	1.2700	1.6164	.78566	.46035	.10379	.20855	.89524
1.07	2.9154	.34301	1.2862	1.6292	.78946	.46470	.10930	.21197	.89733
1.08	2.9447	.33960	1.3025	1.6421	.79320	.46904	.11479	.21541	.89938
1.09	2.9743	.33622	1.3190	1.6552	.79688	.47338	.12025	.21886	.90139
1.10	**3.0042**	**.33287**	**1.3356**	**1.6685**	**.80050**	**0.47772**	**0.12569**	**0.22233**	**$\bar{1}$.90336**
1.11	3.0344	.32956	1.3524	1.6820	.80406	.48207	.13111	.22582	.90529
1.12	3.0649	.32628	1.3693	1.6956	.80757	.48641	.13649	.22931	.90718
1.13	3.0957	.32303	1.3863	1.7093	.81102	.49075	.14186	.23283	.90903
1.14	3.1268	.31982	1.4035	1.7233	.81441	.49510	.14720	.23636	.91085
1.15	3.1582	.31664	1.4208	1.7374	.81775	.49944	.15253	.23990	.91262
1.16	3.1899	.31349	1.4382	1.7517	.82104	.50378	.15783	.24346	.91436
1.17	3.2220	.31037	1.4558	1.7662	.82427	.50812	.16311	.24703	.91607
1.18	3.2544	.30728	1.4735	1.7808	.82745	.51247	.16836	.25062	.91774
1.19	3.2871	.30422	1.4914	1.7957	.83058	.51681	.17360	.25422	.91938
1.20	**3.3201**	**.30119**	**1.5095**	**1.8107**	**.83365**	**0.52115**	**0.17882**	**0.25784**	**$\bar{1}$.92099**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
1.20	**3.3201**	**.30119**	**1.5095**	**1.8107**	**.83365**	**0.52115**	**0.17882**	**0.25784**	**1.92099**
1.21	3.3535	.29820	1.5276	1.8258	.83668	.52550	.18402	.26146	.92256
1.22	3.3872	.29523	1.5460	1.8412	.83965	.52984	.18920	.26510	.92410
1.23	3.4212	.29229	1.5645	1.8568	.84258	.53418	.19437	.26876	.92561
1.24	3.4556	.28938	1.5831	1.8725	.84546	.53853	.19951	.27242	.92709
1.25	3.4903	.28650	1.6019	1.8884	.84828	.54287	.20464	.27610	.92854
1.26	3.5254	.28365	1.6209	1.9045	.85106	.54721	.20975	.27979	.92996
1.27	3.5609	.28083	1.6400	1.9208	.85380	.55155	.21485	.28349	.93135
1.28	3.5966	.27804	1.6593	1.9373	.85648	.55590	.21993	.28721	.93272
1.29	3.6328	.27527	1.6788	1.9540	.85913	.56024	.22499	.29093	.93406
1.30	**3.6693**	**.27253**	**1.6984**	**1.9709**	**.86172**	**0.56458**	**0.23004**	**0.29467**	**$\bar{1}$.93537**
1.31	3.7062	.26982	1.7182	1.9880	.86428	.56893	.23507	.29842	.93665
1.32	3.7434	.26714	1.7381	2.0053	.86678	.57327	.24009	.30217	.93791
1.33	3.7810	.26448	1.7583	2.0228	.86925	.57761	.24509	.30594	.93914
1.34	3.8190	.26185	1.7786	2.0404	.87167	.58195	.25008	.30972	.94035
1.35	3.8574	.25924	1.7991	2.0583	.87405	.58630	.25505	.31352	.94154
1.36	3.8962	.25666	1.8198	2.0764	.87639	.59064	.26002	.31732	.94270
1.37	3.9354	.25411	1.8406	2.0947	.87869	.59498	.26496	.32113	.94384
1.38	3.9749	.25158	1.8617	2.1132	.88095	.59933	.26990	.32495	.94495
1.39	4.0149	.24908	1.8829	2.1320	.88317	.60367	.27482	.32878	.94604
1.40	**4.0552**	**.24660**	**1.9043**	**2.1509**	**.88535**	**0.60801**	**0.27974**	**0.33262**	**$\bar{1}$.94712**
1.41	4.0960	.24414	1.9259	2.1700	.88749	.61236	.28464	.33647	.94817
1.42	4.1371	.24171	1.9477	2.1894	.88960	.61670	.28952	.34033	.94919
1.43	4.1787	.23931	1.9697	2.2090	.89167	.62104	.29440	.34420	.95020
1.44	4.2207	.23693	1.9919	2.2288	.89370	.62538	.29926	.34807	.95119
1.45	4.2631	.23457	2.0143	2.2488	.89569	.62973	.30412	.35196	.95216
1.46	4.3060	.23224	2.0369	2.2691	.89765	.63407	.30896	.35585	.95311
1.47	4.3492	.22993	2.0597	2.2896	.89958	.63841	.31379	.35976	.95404
1.48	4.3929	.22764	2.0827	2.3103	.90147	.64276	.31862	.36367	.95495
1.49	4.4371	.22537	2.1059	2.3312	.90332	.64710	.32343	.36759	.95584
1.50	**4.4817**	**.22313**	**2.1293**	**2.3524**	**.90515**	**0.65144**	**0.32823**	**0.37151**	**$\bar{1}$.95672**
1.51	4.5267	.22091	2.1529	2.3738	.90694	.65578	.33303	.37545	.95758
1.52	4.5722	.21871	2.1768	2.3955	.90870	.66013	.33781	.37939	.95842
1.53	4.6182	.21654	2.2008	2.4174	.91042	.66447	.34258	.38334	.95924
1.54	4.6646	.21438	2.2251	2.4395	.91212	.66881	.34735	.38730	.96005
1.55	4.7115	.21225	2.2496	2.4619	.91379	.67316	.35211	.39126	.96084
1.56	4.7588	.21014	2.2743	2.4845	.91542	.67750	.35686	.39524	.96162
1.57	4.8066	.20805	2.2993	2.5073	.91703	.68184	.36160	.39921	.96238
1.58	4.8550	.20598	2.3245	2.5305	.91860	.68619	.36633	.40320	.96313
1.59	4.9037	.20393	2.3499	2.5538	.92015	.69053	.37105	.40719	.96386
1.60	**4.9530**	**.20190**	**2.3756**	**2.5775**	**.92167**	**0.69487**	**0.37577**	**0.41119**	**$\bar{1}$.96457**
1.61	5.0028	.19989	2.4015	2.6013	.92316	.69921	.38048	.41520	.96528
1.62	5.0531	.19790	2.4276	2.6255	.92462	.70356	.38518	.41921	.96597
1.63	5.1039	.19593	2.4540	2.6499	.92606	.70790	.38987	.42323	.96664
1.64	5.1552	.19398	2.4806	2.6746	.92747	.71224	.39456	.42725	.96730
1.65	5.2070	.19205	2.5075	2.6995	.92886	.71659	.39923	.43129	.96795
1.66	5.2593	.19014	2.5346	2.7247	.93022	.72093	.40391	.43532	.96858
1.67	5.3122	.18825	2.5620	2.7502	.93155	.72527	.40857	.43937	.96921
1.68	5.3656	.18637	2.5896	2.7760	.93286	.72961	.41323	.44341	.96982
1.69	5.4195	.18452	2.6175	2.8020	.93415	.73396	.41788	.44747	.97042
1.70	**5.4739**	**.18268**	**2.6456**	**2.8283**	**.93541**	**0.73830**	**0.42253**	**0.45153**	**$\bar{1}$.97100**
1.71	5.5290	.18087	2.6740	2.8549	.93665	.74264	.42717	.45559	.97158
1.72	5.5845	.17907	2.7027	2.8818	.93786	.74699	.43180	.45966	.97214
1.73	5.6407	.17728	2.7317	2.9090	.93906	.75133	.43643	.46374	.97269
1.74	5.6973	.17552	2.7609	2.9364	.94023	.75567	.44105	.46782	.97323
1.75	5.7546	.17377	2.7904	2.9642	.94138	.76002	.44567	.47191	.97376
1.76	5.8124	.17204	2.8202	2.9922	.94250	.76436	.45028	.47600	.97428
1.77	5.8709	.17033	2.8503	3.0206	.94361	.76870	.45488	.48009	.97479
1.78	5.9299	.16864	2.8806	3.0492	.94470	.77304	.45948	.48419	.97529
1.79	5.9895	.16696	2.9112	3.0782	.94576	.77739	.46408	.48830	.97578
1.80	**6.0496**	**.16530**	**2.9422**	**3.1075**	**.94681**	**0.78173**	**0.46867**	**0.49241**	**$\bar{1}$.97626**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
1.80	**6.0496**	**.16530**	**2.9422**	**3.1075**	**.94681**	**0.78173**	**0.46867**	**0.49241**	**$\bar{1}$.97626**
1.81	6.1104	.16365	2.9734	3.1371	.94783	.78607	.47325	.49652	.97673
1.82	6.1719	.16203	3.0049	3.1669	.94884	.79042	.47783	.50064	.97719
1.83	6.2339	.16041	3.0367	3.1972	.94983	.79476	.48241	.50476	.97764
1.84	6.2965	.15882	3.0689	3.2277	.95080	.79910	.48698	.50889	.97809
1.85	6.3598	.15724	3.1013	3.2585	.95175	.80344	.49154	.51302	.97852
1.86	6.4237	.15567	3.1340	3.2897	.95268	.80779	.49610	.51716	.97895
1.87	6.4883	.15412	3.1671	3.3212	.95359	.81213	.50066	.52130	.97936
1.88	6.5535	.15259	3.2005	3.3530	.95449	.81647	.50521	.52544	.97977
1.89	6.6194	.15107	3.2341	3.3852	.95537	.82082	.50976	.52959	.98017
1.90	**6.6859**	**.14957**	**3.2682**	**3.4177**	**.95624**	**0.82516**	**0.51430**	**0.53374**	**$\bar{1}$.98057**
1.91	6.7531	.14808	3.3025	3.4506	.95709	.82950	.51884	.53789	.98095
1.92	6.8210	.14661	3.3372	3.4838	.95792	.83385	.52338	.54205	.98133
1.93	6.8895	.14515	3.3722	3.5173	.95873	.83819	.52791	.54621	.98170
1.94	6.9588	.14370	3.4075	3.5512	.95953	.84253	.53244	.55038	.98206
1.95	7.0287	.14227	3.4432	3.5855	.96032	.84687	.53696	.55455	.98242
1.96	7.0993	.14086	3.4792	3.6201	.96109	.85122	.54148	.55872	.98272
1.97	7.1707	.13946	3.5156	3.6551	.96185	.85556	.54600	.56290	.98311
1.98	7.2427	.13807	3.5523	3.6904	.96259	.85990	.55051	.56707	.98344
1.99	7.3155	.13670	3.5894	3.7261	.96331	.86425	.55502	.57126	.98377
2.00	**7.3891**	**.13534**	**3.6269**	**3.7622**	**.96403**	**0.86859**	**0.55953**	**0.57544**	**$\bar{1}$.98409**
2.01	7.4633	.13399	3.6647	3.7987	.96473	.87293	.56403	.57963	.98440
2.02	7.5383	.13266	3.7028	3.8355	.96541	.87727	.56853	.58382	.98471
2.03	7.6141	.13134	3.7414	3.8727	.96609	.88162	.57303	.58802	.98502
2.04	7.6906	.13003	3.7803	3.9103	.96675	.88596	.57753	.59221	.98531
2.05	7.7679	.12873	3.8196	3.9483	.96740	.89030	.58202	.59641	.98560
2.06	7.8460	.12745	3.8593	3.9867	.96803	.89465	.58650	.60061	.98589
2.07	7.9248	.12619	3.8993	4.0255	.96865	.89899	.59099	.60482	.98617
2.08	8.0045	.12493	3.9398	4.0647	.96926	.90333	.59547	.60903	.98644
2.09	8.0849	.12369	3.9806	4.1043	.96986	.90768	.59995	.61324	.98671
2.10	**8.1662**	**.12246**	**4.0219**	**4.1443**	**.97045**	**0.91202**	**0.60443**	**0.61745**	**$\bar{1}$.98697**
2.11	8.2482	.12124	4.0635	4.1847	.97103	.91636	.60890	.62167	.98723
2.12	8.3311	.12003	4.1056	4.2256	.97159	.92070	.61337	.62589	.98748
2.13	8.4149	.11884	4.1480	4.2669	.97215	.92505	.61784	.63011	.98773
2.14	8.4994	.11765	4.1909	4.3085	.97269	.92939	.62231	.63433	.98798
2.15	8.5849	.11648	4.2342	4.3507	.97323	.93373	.62677	.63856	.98821
2.16	8.6711	.11533	4.2779	4.3932	.97375	.93808	.63123	.64278	.98845
2.17	8.7583	.11418	4.3221	4.4362	.97426	.94242	.63569	.64701	.98868
2.18	8.8463	.11304	4.3666	4.4797	.97477	.94676	.64015	.65125	.98890
2.19	8.9352	.11192	4.4116	4.5236	.97526	.95110	.64460	.65548	.98912
2.20	**9.0250**	**.11080**	**4.4571**	**4.5679**	**.97574**	**0.95545**	**0.64905**	**0.65972**	**$\bar{1}$.98934**
2.21	9.1157	.10970	4.5030	4.6127	.97622	.95979	.65350	.66396	.98955
2.22	9.2073	.10861	4.5494	4.6580	.97668	.96413	.65795	.66820	.98975
2.23	9.2999	.10753	4.5962	4.7037	.97714	.96848	.66240	.67244	.98996
2.24	9.3933	.10646	4.6434	4.7499	.97759	.97282	.66684	.67668	.99016
2.25	9.4877	.10540	4.6912	4.7966	.97803	.97716	.67128	.68093	.99035
2.26	9.5831	.10435	4.7394	4.8437	.97846	.98151	.67572	.68518	.99054
2.27	9.6794	.10331	4.7880	4.8914	.97888	.98585	.68016	.68943	.99073
2.28	9.7767	.10228	4.8372	4.9395	.97929	.99019	.68459	.69368	.99091
2.29	9.8749	.10127	4.8868	4.9881	.97970	.99453	.68903	.69794	.99109
2.30	**9.9742**	**.10026**	**4.9370**	**5.0372**	**.98010**	**0.99888**	**0.69346**	**0.70219**	**$\bar{1}$.99127**
2.31	10.074	.09926	4.9876	5.0868	.98049	1.00322	.69789	.70645	.99144
2.32	10.176	.09827	5.0387	5.1370	.98087	.00756	.70232	.71071	.99161
2.33	10.278	.09730	5.0903	5.1876	.98124	.01191	.70675	.71497	.99178
2.34	10.381	.09633	5.1425	5.2388	.98161	.01625	.71117	.71923	.99194
2.35	10.486	.09537	5.1951	5.2905	.98197	.02059	.71559	.72349	.99210
2.36	10.591	.09442	5.2483	5.3427	.98233	.02493	.72002	.72776	.99226
2.37	10.697	.09348	5.3020	5.3954	.98267	.02928	.72444	.73203	.99241
2.38	10.805	.09255	5.3562	5.4487	.98301	.03362	.72885	.73630	.99256
2.39	10.913	.09163	5.4109	5.5026	.98335	.03796	.73327	.74056	.99271
2.40	**11.023**	**.09072**	**5.4662**	**5.5569**	**.98367**	**1.04231**	**0.73769**	**0.74484**	**$\bar{1}$.99285**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
2.40	**11.023**	**.09072**	**5.4662**	**5.5569**	**.98367**	**1.04231**	**0.73769**	**0.74484**	**$\bar{1}$.99285**
2.41	11.134	.08982	5.5221	5.6119	.98400	.04665	.74210	.74911	.99299
2.42	11.246	.08892	5.5785	5.6674	.98431	.05099	.74652	.75338	.99313
2.43	11.359	.08804	5.6354	5.7235	.98462	.05534	.75093	.75766	.99327
2.44	11.473	.08716	5.6929	5.7801	.98492	.05968	.75534	.76194	.99340
2.45	11.588	.08629	5.7510	5.8373	.98522	.06402	.75975	.76621	.99353
2.46	11.705	.08543	5.8097	5.8951	.98551	.06836	.76415	.77049	.99366
2.47	11.822	.08458	5.8689	5.9535	.98579	.07271	.76856	.77477	.99379
2.48	11.941	.08374	5.9288	6.0125	.98607	.07705	.77296	.77906	.99391
2.49	12.061	.08291	5.9892	6.0721	.98635	.08139	.77737	.78334	.99403
2.50	**12.182**	**.08208**	**6.0502**	**6.1323**	**.98661**	**1.08574**	**0.78177**	**0.78762**	**$\bar{1}$.99415**
2.51	12.305	.08127	6.1118	6.1931	.98688	.09008	.78617	.79191	.99426
2.52	12.429	.08046	6.1741	6.2545	.98714	.09442	.79057	.79619	.99438
2.53	12.554	.07966	6.2369	6.3166	.98739	.09877	.79497	.80048	.99449
2.54	12.680	.07887	6.3004	6.3793	.98764	.10311	.79937	.80477	.99460
2.55	12.807	.07808	6.3645	6.4426	.98788	.10745	.80377	.80906	.99470
2.56	12.936	.07730	6.4293	6.5066	.98812	.11179	.80816	.81335	.99481
2.57	13.066	.07654	6.4946	6.5712	.98835	.11614	.81256	.81764	.99491
2.58	13.197	.07577	6.5607	6.6365	.98858	.12048	.81695	.82194	.99501
2.59	13.330	.07502	6.6274	6.7024	.98881	.12482	.82134	.82623	.99511
2.60	**13.464**	**.07427**	**6.6947**	**6.7690**	**.98903**	**1.12917**	**0.82573**	**0.83052**	**$\bar{1}$.99521**
2.61	13.599	.07353	6.7628	6.8363	.98924	.13351	.83012	.83482	.99530
2.62	13.736	.07280	6.8315	6.9043	.98946	.13785	.83451	.83912	.99540
2.63	13.874	.07208	6.9008	6.9729	.98966	.14219	.83890	.84341	.99549
2.64	14.013	.07136	6.9709	7.0423	.98987	.14654	.84329	.84771	.99558
2.65	14.154	.07065	7.0417	7.1123	.99007	.15088	.84768	.85201	.99566
2.66	14.296	.06995	7.1132	7.1831	.99026	.15522	.85206	.85631	.99575
2.67	14.440	.06925	7.1854	7.2546	.99045	.15957	.85645	.86061	.99583
2.68	14.585	.06856	7.2583	7.3268	.99064	.16391	.86083	.86492	.99592
2.69	14.732	.06788	7.3319	7.3998	.99083	.16825	.86522	.86922	.99600
2.70	**14.880**	**.06721**	**7.4063**	**7.4735**	**.99101**	**1.17260**	**0.86960**	**0.87352**	**$\bar{1}$.99608**
2.71	15.029	.06654	7.4814	7.5479	.99118	.17694	.87398	.87783	.99615
2.72	15.180	.06587	7.5572	7.6231	.99136	.18128	.87836	.88213	.99623
2.73	15.333	.06522	7.6338	7.6991	.99153	.18562	.88274	.88644	.99631
2.74	15.487	.06457	7.7112	7.7758	.99170	.18997	.88712	.89074	.99638
2.75	15.643	.06393	7.7894	7.8533	.99186	.19431	.89150	.89505	.99645
2.76	15.800	.06329	7.8683	7.9316	.99202	.19865	.89588	.89936	.99652
2.77	15.959	.06266	7.9480	8.0106	.99218	.20300	.90026	.90367	.99659
2.78	16.119	.06204	8.0285	8.0905	.99233	.20734	.90463	.90798	.99666
2.79	16.281	.06142	8.1098	8.1712	.99248	.21168	.90901	.91229	.99672
2.80	**16.445**	**.06081**	**8.1919**	**8.2527**	**.99263**	**1.21602**	**0.91339**	**0.91660**	**$\bar{1}$.99679**
2.81	16.610	.06020	8.2749	8.3351	.99278	.22037	.91776	.92091	.99685
2.82	16.777	.05961	8.3586	8.4182	.99292	.22471	.92213	.92522	.99691
2.83	16.945	.05901	8.4432	8.5022	.99306	.22905	.92651	.92953	.99698
2.84	17.116	.05843	8.5287	8.5871	.99320	.23340	.93088	.93385	.99704
2.85	17.288	.05784	8.6150	8.6728	.99333	.23774	.93525	.93816	.99709
2.86	17.462	.05727	8.7021	8.7594	.99346	.24208	.93963	.94247	.99715
2.87	17.637	.05670	8.7902	8.8469	.99359	.24643	.94400	.94679	.99721
2.88	17.814	.05613	8.8791	8.9352	.99372	.25077	.94837	.95110	.99726
2.89	17.993	.05558	8.9689	9.0244	.99384	.25511	.95274	.95542	.99732
2.90	**18.174**	**.05502**	**9.0596**	**9.1146**	**.99396**	**1.25945**	**0.95711**	**0.95974**	**$\bar{1}$.99737**
2.91	18.357	.05448	9.1512	9.2056	.99408	.26380	.96148	.96405	.99742
2.92	18.541	.05393	9.2437	9.2976	.99420	.26814	.96584	.96837	.99747
2.93	18.728	.05340	9.3371	9.3905	.99431	.27248	.97021	.97269	.99752
2.94	18.916	.05287	9.4315	9.4844	.99443	.27683	.97458	.97701	.99757
2.95	19.106	.05234	9.5268	9.5791	.99454	.28117	.97895	.98133	.99762
2.96	19.298	.05182	9.6231	9.6749	.99464	.28551	.98331	.98565	.99767
2.97	19.492	.05130	9.7203	9.7716	.99475	.28985	.98768	.98997	.99771
2.98	19.688	.05079	9.8185	9.8693	.99485	.29420	.99205	.99429	.99776
2.99	19.886	.05029	9.9177	9.9680	.99496	.29854	.99641	.99861	.99780
3.00	**20.086**	**.04979**	**10.018**	**10.068**	**.99505**	**1.30288**	**1.00078**	**1.00293**	**$\bar{1}$.99785**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
3.00	**20.086**	**.04979**	**10.018**	**10.068**	**.99505**	**1.30288**	**1.00078**	**1.00293**	**$\bar{1}$.99785**
3.01	20.287	.04929	10.119	10.168	.99515	.30723	.00514	.00725	.99789
3.02	20.491	.04880	10.221	10.270	.99525	.31157	.00950	.01157	.99793
3.03	20.697	.04832	10.325	10.373	.99534	.31591	.01387	.01589	.99797
3.04	20.905	.04783	10.429	10.477	.99543	.32026	.01823	.02022	.99801
3.05	21.115	.04736	10.534	10.581	.99552	.32460	.02259	.02454	.99805
3.06	21.328	.04689	10.640	10.687	.99561	.32894	.02696	.02886	.99809
3.07	21.542	.04642	10.748	10.794	.99570	.33328	.03132	.03319	.99813
3.08	21.758	.04596	10.856	10.902	.99578	.33763	.03568	.03751	.99817
3.09	21.977	.04550	10.966	11.011	.99587	.34197	.04004	.04184	.99820
3.10	**22.198**	**.04505**	**11.077**	**11.122**	**.99595**	**1.34631**	**1.04440**	**1.04616**	**$\bar{1}$.99824**
3.11	22.421	.04460	11.188	11.233	.99603	.35066	.04876	.05049	.99827
3.12	22.646	.04416	11.301	11.345	.99611	.35500	.05312	.05481	.99831
3.13	22.874	.04372	11.415	11.459	.99618	.35934	.05748	.05914	.99834
3.14	23.104	.04328	11.530	11.574	.99626	.36368	.06184	.06347	.99837
3.15	23.336	.04285	11.647	11.689	.99633	.36803	.06620	.06779	.99841
3.16	23.571	.04243	11.764	11.807	.99641	.37237	.07056	.07212	.99844
3.17	23.807	.04200	11.883	11.925	.99648	.37671	.07492	.07645	.99847
3.18	24.047	.04159	12.003	12.044	.99655	.38106	.07927	.08078	.99850
3.19	24.288	.04117	12.124	12.165	.99662	.38540	.08363	.08510	.99853
3.20	**24.533**	**.04076**	**12.246**	**12.287**	**.99668**	**1.38974**	**1.08799**	**1.08943**	**$\bar{1}$.99856**
3.21	24.779	.04036	12.369	12.410	.99675	.39409	.09235	.09376	.99859
3.22	25.028	.03996	12.494	12.534	.99681	.39843	.09670	.09809	.99861
3.23	25.280	.03956	12.620	12.660	.99688	.40277	.10106	.10242	.99864
3.24	25.534	.03916	12.747	12.786	.99694	.40711	.10542	.10675	.99867
3.25	25.790	.03877	12.876	12.915	.99700	.41146	.10977	.11108	.99869
3.26	26.050	.03839	13.006	13.044	.99706	.41580	.11413	.11541	.99872
3.27	26.311	.03801	13.137	13.175	.99712	.42014	.11849	.11974	.99875
3.28	26.576	.03763	13.269	13.307	.99717	.42449	.12284	.12407	.99877
3.29	26.843	.03725	13.403	13.440	.99723	.42883	.12720	.12840	.99879
3.30	**27.113**	**.03688**	**13.538**	**13.575**	**.99728**	**1.43317**	**1.13155**	**1.13273**	**$\bar{1}$.99882**
3.31	27.385	.03652	13.674	13.711	.99734	43751	.13591	.13706	.99884
3.32	27.660	.03615	13.812	13.848	.99739	.44186	.14026	.14139	.99886
3.33	27.938	.03579	13.951	13.987	.99744	.44620	.14461	.14573	.99889
3.34	28.219	.03544	14.092	14.127	.99749	.45054	.14897	.15006	.99891
3.35	28.503	.03508	14.234	14.269	.99754	.45489	.15332	.15439	.99893
3.36	28.789	.03474	14.377	14.412	.99759	.45923	.15768	.15872	.99895
3.37	29.079	.03439	14.522	14.556	.99764	.46357	.16203	.16306	.99897
3.38	29.371	.03405	14.668	14.702	.99768	.46792	.16638	.16739	.99899
3.39	29.666	.03371	14.816	14.850	.99773	.47226	.17073	.17172	.99901
3.40	**29.964**	**.03337**	**14.965**	**14.999**	**.99777**	**1.47660**	**1.17509**	**1.17605**	**$\bar{1}$.99903**
3.41	30.265	.03304	15.116	15.149	.99782	.48094	.17944	.18039	.99905
3.42	30.569	.03271	15.268	15.301	.99786	.48529	.18379	.18472	.99907
3.43	30.877	.03239	15.422	15.455	.99790	.48963	.18814	.18906	.99909
3.44	31.187	.03206	15.577	15.610	.99795	.49397	.19250	.19339	.99911
3.45	31.500	.03175	15.734	15.766	.99799	.49832	.19685	.19772	.99912
3.46	31.817	.03143	15.893	15.924	.99803	.50266	.20120	.20206	.99914
3.47	32.137	.03112	16.053	16.084	.99807	.50700	.20555	.20639	.99916
3.48	32.460	.03081	16.215	16.245	.99810	.51134	.20990	.21073	.99918
3.49	32.786	.03050	16.378	16.408	.99814	.51569	.21425	.21506	.99919
3.50	**33.115**	**.03020**	**16.543**	**16.573**	**.99818**	**1.52003**	**1.21860**	**1.21940**	**$\bar{1}$.99921**
3.51	33.448	.02990	16.709	16.739	.99821	.52437	.22296	.22373	.99922
3.52	33.784	.02960	16.877	16.907	.99825	.52872	.22731	.22807	.99924
3.53	34.124	.02930	17.047	17.077	.99828	.53306	.23166	.23240	.99925
3.54	34.467	.02901	17.219	17.248	.99832	.53740	.23601	.23674	.99927
3.55	34.813	.02872	17.392	17.421	.99835	.54175	.24036	.24107	.99928
3.56	35.163	.02844	17.567	17.596	.99838	.54609	.24471	.24541	.99930
3.57	35.517	.02816	17.744	17.772	.99842	.55043	.24906	.24975	.99931
3.58	35.874	.02788	17.923	17.951	.99845	.55477	.25341	.25408	.99933
3.59	36.234	.02760	18.103	18.131	.99848	.55912	.25776	.25842	.99934
3.60	**36.598**	**.02732**	**18.285**	**18.313**	**.99851**	**1.56346**	**1.26211**	**1.26275**	**$\bar{1}$.99935**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
3.60	**36.598**	**.02732**	**18.285**	**18.313**	**.99851**	**1.56346**	**1.26211**	**1.26275**	**$\bar{1}$.99935**
3.61	36.966	.02705	18.470	18.497	.99854	.56780	.26646	.26709	.99936
3.62	37.338	.02678	18.655	18.682	.99857	.57215	.27080	.27143	.99938
3.63	37.713	.02652	18.843	18.870	.99859	.57649	.27515	.27576	.99939
3.64	38.092	.02625	19.033	19.059	.99862	.58083	.27950	.28010	.99940
3.65	38.475	.02599	19.224	19.250	.99865	.58517	.28385	.28444	.99941
3.66	38.861	.02573	19.418	19.444	.99868	.58952	.28820	.28878	.99942
3.67	39.252	.02548	19.613	19.639	.99870	.59386	.29255	.29311	.99944
3.68	39.646	.02522	19.811	19.836	.99873	.59820	.29690	.29745	.99945
3.69	40.045	.02497	20.010	20.035	.99875	.60255	.30125	.30179	.99946
3.70	**40.447**	**.02472**	**20.211**	**20.236**	**.99878**	**1.60689**	**1.30559**	**1.30612**	**$\bar{1}$.99947**
3.71	40.854	.02448	20.415	20.439	.99880	.61123	.30994	.31046	.99948
3.72	41.264	.02423	20.620	20.644	.99883	.61558	.31429	.31480	.99949
3.73	41.679	.02399	20.828	20.852	.99885	.61992	.31864	.31914	.99950
3.74	42.098	.02375	21.037	21.061	.99887	.62426	.32299	.32348	.99951
3.75	42.521	.02352	21.249	21.272	.99889	.62860	.32733	.32781	.99952
3.76	42.948	.02328	21.463	21.486	.99892	.63295	.33168	.33215	.99953
3.77	43.380	.02305	21.679	21.702	.99894	.63729	.33603	.33649	.99954
3.78	43.816	.02282	21.897	21.919	.99896	.64163	.34038	.34083	.99955
3.79	44.256	.02260	22.117	22.140	.99898	.64598	.34472	.34517	.99956
3.80	**44.701**	**.02237**	**22.339**	**22.362**	**.99900**	**1.65032**	**1.34907**	**1.34951**	**$\bar{1}$.99957**
3.81	45.150	.02215	22.564	22.586	.99902	.65466	.35342	.35384	.99957
3.82	45.604	.02193	22.791	22.813	.99904	.65900	.35777	.35818	.99958
3.83	46.063	.02171	23.020	23.042	.99906	.66335	.36211	.36252	.99959
3.84	46.525	.02149	23.252	23.274	.99908	.66769	.36646	.36686	.99960
3.85	46.993	.02128	23.486	23.507	.99909	.67203	.37081	.37120	.99961
3.86	47.465	.02107	23.722	23.743	.99911	.67638	.37515	.37554	.99961
3.87	47.942	.02086	23.961	23.982	.99913	.68072	.37950	.37988	.99962
3.88	48.424	.02065	24.202	24.222	.99915	.68506	.38385	.38422	.99963
3.89	48.911	.02045	24.445	24.466	.99916	.68941	.38819	.38856	.99964
3.90	**49.402**	**.02024**	**24.691**	**24.711**	**.99918**	**1.69375**	**1.39254**	**1.39290**	**$\bar{1}$.99964**
3.91	49.899	.02004	24.939	24.960	.99920	.69809	.39689	.39724	.99965
3.92	50.400	.01984	25.190	25.210	.99921	.70243	.40123	.40158	.99966
3.93	50.907	.01964	25.444	25.463	.99923	.70678	.40558	.40591	.99966
3.94	51.419	.01945	25.700	25.719	.99924	.71112	.40993	.41025	.99967
3.95	51.935	.01925	25.958	25.977	.99926	.71546	.41427	.41459	.99968
3.96	52.457	.01906	26.219	26.238	.99927	.71981	.41862	.41893	.99968
3.97	52.985	.01887	26.483	26.502	.99929	.72415	.42296	.42327	.99969
3.98	53.517	.01869	26.749	26.768	.99930	.72849	.42731	.42761	.99970
3.99	54.055	.01850	27.018	27.037	.99932	.73284	.43166	.43195	.99970
4.00	**54.598**	**.01832**	**27.290**	**27.308**	**.99933**	**1.73718**	**1.43600**	**1.43629**	**$\bar{1}$.99971**
4.01	55.147	.01813	27.564	27.583	.99934	.74152	.44035	.44063	.99971
4.02	55.701	.01795	27.842	27.860	.99936	.74586	.44469	.44497	.99972
4.03	56.261	.01777	28.122	28.139	.99937	.75021	.44904	.44931	.99973
4.04	56.826	.01760	28.404	28.422	.99938	.75455	.45339	.45365	.99973
4.05	57.397	.01742	28.690	28.707	.99939	.75889	.45773	.45799	.99974
4.06	57.974	.01725	28.979	28.996	.99941	.76324	.46208	.46233	.99974
4.07	58.557	.01708	29.270	29.287	.99942	.76758	.46642	.46668	.99975
4.08	59.145	.01691	29.564	29.581	.99943	.77192	.47077	.47102	.99975
4.09	59.740	.01674	29.862	29.878	.99944	.77626	.47511	.47536	.99976
4.10	**60.340**	**.01657**	**30.162**	**30.178**	**.99945**	**1.78061**	**1.47946**	**1.47970**	**$\bar{1}$.99976**
4.11	60.947	.01641	30.465	30.482	.99946	.78495	.48380	.48404	.99977
4.12	61.559	.01624	30.772	30.788	.99947	.78929	.48815	.48838	.99977
4.13	62.178	.01608	31.081	31.097	.99948	.79364	.49249	.49272	.99978
4.14	62.803	.01592	31.393	31.409	.99949	.79798	.49684	.49706	.99978
4.15	63.434	.01576	31.709	31.725	.99950	.80232	.50118	.50140	.99978
4.16	64.072	.01561	32.028	32.044	.99951	.80667	.50553	.50574	.99979
4.17	64.715	.01545	32.350	32.365	.99952	.81101	.50987	.51008	.99979
4.18	65.366	.01530	32.675	32.691	.99953	.81535	.51422	.51442	.99980
4.19	66.023	.01515	33.004	33.019	.99954	.81969	.51856	.51876	.99980
4.20	**66.686**	**.01500**	**33.336**	**33.351**	**.99955**	**1.82404**	**1.52291**	**1.52310**	**$\bar{1}$.99980**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
4.20	**66.686**	**.01500**	**33.336**	**33.351**	**.99955**	**1.82404**	**1.52291**	**1.52310**	**$\bar{1}$.99980**
4.21	67.357	.01485	33.671	33.686	.99956	.82838	.52725	.52745	.99981
4.22	68.033	.01470	34.009	34.024	.99957	.83272	.53160	.53179	.99981
4.23	68.717	.01455	34.351	34.366	.99958	.83707	.53594	.53613	.99982
4.24	69.408	.01441	34.697	34.711	.99958	.84141	.54029	.54047	.99982
4.25	70.105	.01426	35.046	35.060	.99959	.84575	.54463	.54481	.99982
4.26	70.810	.01412	35.398	35.412	.99960	.85009	.54898	.54915	.99983
4.27	71.522	.01398	35.754	35.768	.99961	.85444	.55332	.55349	.99983
4.28	72.240	.01384	36.113	36.127	.99962	.85878	.55767	.55783	.99983
4.29	72.966	.01370	36.476	36.490	.99962	.86312	.56201	.56217	.99984
4.30	**73.700**	**.01357**	**36.843**	**36.857**	**.99963**	**1.86747**	**1.56636**	**1.56652**	**$\bar{1}$.99984**
4.31	74.440	.01343	37.214	37.227	.99964	.87181	.57070	.57086	.99984
4.32	75.189	.01330	37.588	37.601	.99965	.87615	.57505	.57520	.99985
4.33	75.944	.01317	37.966	37.979	.99965	.88050	.57939	.57954	.99985
4.34	76.708	.01304	38.347	38.360	.99966	.88484	.58373	.58388	.99985
4.35	77.478	.01291	38.733	38.746	.99967	.88918	.58808	.58822	.99986
4.36	78.257	.01278	39.122	39.135	.99967	.89352	.59242	.59256	.99986
4.37	79.044	.01265	39.515	39.528	.99968	.89787	.59677	.59691	.99986
4.38	79.838	.01253	39.913	39.925	.99969	.90221	.60111	.60125	.99986
4.39	80.640	.01240	40.314	40.326	.99969	.90655	.60546	.60559	.99987
4.40	**81.451**	**.01228**	**40.719**	**40.732**	**.99970**	**1.91090**	**1.60980**	**1.60993**	**$\bar{1}$.99987**
4.41	82.269	.01216	41.129	41.141	.99970	.91524	.61414	.61427	.99987
4.42	83.096	.01203	41.542	41.554	.99971	.91958	.61849	.61861	.99987
4.43	83.931	.01191	41.960	41.972	.99972	.92392	.62283	.62296	.99988
4.44	84.775	.01180	42.382	42.393	.99972	.92827	.62718	.62730	.99988
4.45	85.627	.01168	42.808	42.819	.99973	.93261	.63152	.63164	.99988
4.46	86.488	.01156	43.238	43.250	.99973	.93695	.63587	.63598	.99988
4.47	87.357	.01145	43.673	43.684	.99974	.94130	.64021	.64032	.99989
4.48	88.235	.01133	44.112	44.123	.99974	.94564	.64455	.64467	.99989
4.49	89.121	.01122	44.555	44.566	.99975	.94998	.64890	.64901	.99989
4.50	**90.017**	**.01111**	**45.003**	**45.014**	**.99975**	**1.95433**	**1.65324**	**1.65335**	**$\bar{1}$.99989**
4.51	90.922	.01100	45.455	45.466	.99976	.95867	.65759	.65769	.99989
4.52	91.836	.01089	45.912	45.923	.99976	.96301	.66193	.66203	.99990
4.53	92.759	.01078	46.374	46.385	.99977	.96735	.66627	.66637	.99990
4.54	93.691	.01067	46.840	46.851	.99977	.97170	.67062	.67072	.99990
4.55	94.632	.01057	47.311	47.321	.99978	.97604	.67496	.67506	.99990
4.56	95.583	.01046	47.787	47.797	.99978	.98038	.67931	.67940	.99990
4.57	96.544	.01036	48.267	48.277	.99979	.98473	.68365	.68374	.99991
4.58	97.514	.01025	48.752	48.762	.99979	.98907	.68799	.68808	.99991
4.59	98.494	.01015	49.242	49.252	.99979	.99341	.69234	.69243	.99991
4.60	**99.484**	**.01005**	**49.737**	**49.747**	**.99980**	**1.99775**	**1.69668**	**1.69677**	**$\bar{1}$.99991**
4.61	100.48	.00995	50.237	50.247	.99980	2.00210	.70102	.70111	.99991
4.62	101.49	.00985	50.742	50.752	.99981	.00644	.70537	.70545	.99992
4.63	102.51	.00975	51.252	51.262	.99981	.01078	.70971	.70979	.99992
4.64	103.54	.00966	51.767	51.777	.99981	.01513	.71406	.71414	.99992
4.65	104.58	.00956	52.288	52.297	.99982	.01947	.71840	.71848	.99992
4.66	105.64	.00947	52.813	52.823	.99982	.02381	.72274	.72282	.99992
4.67	106.70	.00937	53.344	53.354	.99982	.02816	.72709	.72716	.99992
4.68	107.77	.00928	53.880	53.890	.99983	.03250	.73143	.73151	.99993
4.69	108.85	.00919	54.422	54.431	.99983	.03684	.73577	.73585	.99993
4.70	**109.95**	**.00910**	**54.969**	**54.978**	**.99983**	**2.04118**	**1.74012**	**1.74019**	**$\bar{1}$.99993**
4.71	111.05	.00900	55.522	55.531	.99984	.04553	.74446	.74453	.99993
4.72	112.17	.00892	56.080	56.089	.99984	.04987	.74881	.74887	.99993
4.73	113.30	.00883	56.643	56.652	.99984	.05421	.75315	.75322	.99993
4.74	114.43	.00874	57.213	57.222	.99985	.05856	.75749	.75756	.99993
4.75	115.58	.00865	57.788	57.796	.99985	.06290	.76184	.76190	.99993
4.76	116.75	.00857	58.369	58.377	.99985	.06724	.76618	.76624	.99994
4.77	117.92	.00848	58.955	58.964	.99986	.07158	.77052	.77059	.99994
4.78	119.10	.00840	59.548	59.556	.99986	.07593	.77487	.77493	.99994
4.79	120.30	.00831	60.147	60.155	.99986	.08027	.77921	.77927	.99994
4.80	**121.51**	**.00823**	**60.751**	**60.759**	**.99986**	**2.08461**	**1.78355**	**1.78361**	**$\bar{1}$.99994**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
4.80	**121.51**	**.00823**	**60.751**	**60.760**	**.99986**	**2.08461**	**1.78355**	**1.78361**	**$\bar{1}$.99994**
4.81	122.73	.00815	61.362	61.370	.99987	.08896	.78790	.78796	.99994
4.82	123.97	.00807	61.979	61.987	.99987	.09330	.79224	.79230	.99994
4.83	125.21	.00799	62.601	62.609	.99987	.09764	.79658	.79664	.99994
4.84	126.47	.00791	63.231	63.239	.99987	.10199	.80093	.80098	.99995
4.85	127.74	.00783	63.866	63.874	.99988	.10633	.80527	.80532	.99995
4.86	129.02	.00775	64.508	64.516	.99988	.11067	.80962	.80967	.99995
4.87	130.32	.00767	65.157	65.164	.99988	.11501	.81396	.81401	.99995
4.88	131.63	.00760	65.812	65.819	.99988	.11936	.81830	.81835	.99995
4.89	132.95	.00752	66.473	66.481	.99989	.12370	.82265	.82269	.99995
4.90	**134.29**	**.00745**	**67.141**	**67.149**	**.99989**	**2.12804**	**1.82699**	**1.82704**	**$\bar{1}$.99995**
4.91	135.64	.00737	67.816	67.823	.99989	.13239	.83133	.83138	.99995
4.92	137.00	.00730	68.498	68.505	.99989	.13673	.83568	.83572	.99995
4.93	138.38	.00723	69.186	69.193	.99990	.14107	.84002	.84006	.99995
4.94	139.77	.00715	69.882	69.889	.99990	.14541	.84436	.84441	.99996
4.95	141.17	.00708	70.584	70.591	.99990	.14976	.84871	.84875	.99996
4.96	142.59	.00701	71.293	71.300	.99990	.15410	.85305	.85309	.99996
4.97	144.03	.00694	72.010	72.017	.99990	.15844	.85739	.85743	.99996
4.98	145.47	.00687	72.734	72.741	.99991	.16279	.86174	.86178	.99996
4.99	146.94	.00681	73.465	73.472	.99991	.16713	.86608	.86612	.99996
5.00	**148.41**	**.00674**	**74.203**	**74.210**	**.99991**	**2.17147**	**1.87042**	**1.87046**	**$\bar{1}$.99996**
5.01	149.90	.00667	74.949	74.956	.99991	.17582	.87477	.87480	.99996
5.02	151.41	.00660	75.702	75.710	.99991	.18016	.87911	.87915	.99996
5.03	152.93	.00654	76.463	76.470	.99991	.18450	.88345	.88349	.99996
5.04	154.47	.00647	77.232	77.238	.99992	.18884	.88780	.88783	.99996
5.05	156.02	.00641	78.008	78.014	.99992	.19319	.89214	.89217	.99996
5.06	157.59	.00635	78.792	78.798	.99992	.19753	.89648	.89652	.99997
5.07	159.17	.00628	79.584	79.590	.99992	.20187	.90083	.90086	.99997
5.08	160.77	.00622	80.384	80.390	.99992	.20622	.90517	.90520	.99997
5.09	162.39	.00616	81.192	81.198	.99992	.21056	.90951	.90955	.99997
5.10	**164.02**	**.00610**	**82.008**	**82.014**	**.99993**	**2.21490**	**1.91386**	**1.91389**	**$\bar{1}$.99997**
5.11	165.67	.00604	82.832	82.838	.99993	.21924	.91820	.91823	.99997
5.12	167.34	.00598	83.665	83.671	.99993	.22359	.92254	.92257	.99997
5.13	169.02	.00592	84.506	84.512	.99993	.22793	.92689	.92692	.99997
5.14	170.72	.00586	85.355	85.361	.99993	.23227	.93123	.93126	.99997
5.15	172.43	.00580	86.213	86.219	.99993	.23662	.93557	.93560	.99997
5.16	174.16	.00574	87.079	87.085	.99993	.24096	.93992	.93994	.99997
5.17	175.91	.00568	87.955	87.960	.99994	.24530	.94426	.94429	.99997
5.18	177.68	.00563	88.839	88.844	.99994	.24965	.94860	.94863	.99997
5.19	179.47	.00557	89.732	89.737	.99994	.25399	.95294	.95297	.99997
5.20	**181.27**	**.00552**	**90.633**	**90.639**	**.99994**	**2.25833**	**1.95729**	**1.95731**	**$\bar{1}$.99997**
5.21	183.09	.00546	91.544	91.550	.99994	.26267	.96163	.96166	.99997
5.22	184.93	.00541	92.464	92.470	.99994	.26702	.96597	.96600	.99997
5.23	186.79	.00535	93.394	93.399	.99994	.27136	.97032	.97034	.99998
5.24	188.67	.00530	94.332	94.338	.99994	.27570	.97466	.97469	.99998
5.25	190.57	.00525	95.281	95.286	.99994	.28005	.97900	.97903	.99998
5.26	192.48	.00520	96.238	96.243	.99995	.28439	.98335	.98337	.99998
5.27	194.42	.00514	97.205	97.211	.99995	.28873	.98769	.98771	.99998
5.28	196.37	.00509	98.182	98.188	.99995	.29307	.99203	.99206	.99998
5.29	198.34	.00504	99.169	99.174	.99995	.29742	.99638	.99640	.99998
5.30	**200.34**	**.00499**	**100.17**	**100.17**	**.99995**	**2.30176**	**2.00072**	**2.00074**	**$\bar{1}$.99998**
5.31	202.35	.00494	101.17	101.18	.99995	.30610	.00506	.00508	.99998
5.32	204.38	.00489	102.19	102.19	.99995	.31045	.00941	.00943	.99998
5.33	206.44	.00484	103.22	103.22	.99995	.31479	.01375	.01377	.99998
5.34	208.51	.00480	104.25	104.26	.99995	.31913	.01809	.01811	.99998
5.35	210.61	.00475	105.30	105.31	.99995	.32348	.02244	.02246	.99998
5.36	212.72	.00470	106.36	106.36	.99996	.32782	.02678	.02680	.99998
5.37	214.86	.00465	107.43	107.43	.99996	.33216	.03112	.03114	.99998
5.38	217.02	.00461	108.51	108.51	.99996	.33650	.03547	.03548	.99998
5.39	219.20	.00456	109.60	109.60	.99996	.34085	.03981	.03983	.99998
5.40	**221.41**	**.00452**	**110.70**	**110.71**	**.99996**	**2.34519**	**2.04415**	**2.04417**	**$\bar{1}$.99998**

x	Natural Values					Common Logarithms			
	e^x	e^{-x}	Sinh x	Cosh x	Tanh x	e^x	Sinh x	Cosh x	Tanh x
5.40	**221.41**	**.00452**	**110.70**	**110.71**	**.99996**	**2.34519**	**2.04415**	**2.04417**	**$\bar{1}$.99998**
5.41	223.63	.00447	111.81	111.82	.99996	.34953	.04849	.04851	.99998
5.42	225.88	.00443	112.94	112.94	.99996	.35388	.05284	.05285	.99998
5.43	228.15	.00438	114.07	114.08	.99996	.35822	.05718	.05720	.99998
5.44	230.44	.00434	115.22	115.22	.99996	.36256	.06152	.06154	.99998
5.45	232.76	.00430	116.38	116.38	.99996	.36690	.06587	.06588	.99998
5.46	235.10	.00425	117.55	117.55	.99996	.37125	.07021	.07023	.99998
5.47	237.46	.00421	118.73	118.73	.99996	.37559	.07455	.07457	.99998
5.48	239.85	.00417	119.92	119.93	.99997	.37993	.07890	.07891	.99998
5.49	242.26	.00413	121.13	121.13	.99997	.38428	.08324	.08325	.99999
5.50	**244.69**	**.00409**	**122.34**	**122.35**	**.99997**	**2.38862**	**2.08758**	**2.08760**	**$\bar{1}$.99999**
5.51	247.15	.00405	123.57	123.58	.99997	.39296	.09193	.09194	.99999
5.52	249.64	.00401	124.82	124.82	.99997	.39731	.09627	.09628	.99999
5.53	252.14	.00397	126.07	126.07	.99997	.40165	.10061	.10063	.99999
5.54	254.68	.00393	127.34	127.34	.99997	.40599	.10495	.10497	.99999
5.55	257.24	.00389	128.62	128.62	.99997	.41033	.10930	.10931	.99999
5.56	259.82	.00385	129.91	129.91	.99997	.41468	.11364	.11365	.99999
5.57	262.43	.00381	131.22	131.22	.99997	.41902	.11798	.11800	.99999
5.58	265.07	.00377	132.53	132.54	.99997	.42336	.12233	.12234	.99999
5.59	267.74	.00374	133.87	133.87	.99997	.42771	.12667	.12668	.99999
5.60	**270.43**	**.00370**	**135.21**	**135.22**	**.99997**	**2.43205**	**2.13101**	**2.13103**	**$\bar{1}$.99999**
5.61	273.14	.00366	136.57	136.57	.99997	.43639	.13536	.13537	.99999
5.62	275.89	.00362	137.94	137.95	.99997	.44074	.13970	.13971	.99999
5.63	278.66	.00359	139.33	139.33	.99997	.44508	.14404	.14405	.99999
5.64	281.46	.00355	140.73	140.73	.99997	.44942	.14839	.14840	.99999
5.65	284.29	.00352	142.14	142.15	.99998	.45376	.15273	.15274	.99999
5.66	287.15	.00348	143.57	143.58	.99998	.45811	.15707	.15708	.99999
5.67	290.03	.00345	145.02	145.02	.99998	.46245	.16141	.16142	.99999
5.68	292.95	.00341	146.47	146.48	.99998	.46679	.16576	.16577	.99999
5.69	295.89	.00338	147.95	147.95	.99998	.47114	.17010	.17011	.99999
5.70	**298.87**	**.00335**	**149.43**	**149.44**	**.99998**	**2.47548**	**2.17444**	**2.17445**	**$\bar{1}$.99999**
5.71	301.87	.00331	150.93	150.94	.99998	.47982	.17879	.17880	.99999
5.72	304.90	.00328	152.45	152.45	.99998	.48416	.18313	.18314	.99999
5.73	307.97	.00325	153.98	153.99	.99998	.48851	.18747	.18748	.99999
5.74	311.06	.00321	155.53	155.53	.99998	.49285	.19182	.19182	.99999
5.75	314.19	.00318	157.09	157.10	.99998	.49719	.19616	.19617	.99999
5.76	317.35	.00315	158.67	158.68	.99998	.50154	.20050	.20051	.99999
5.77	320.54	.00312	160.27	160.27	.99998	.50588	.20484	.20485	.99999
5.78	323.76	.00309	161.88	161.88	.99998	.51022	.20919	.20920	.99999
5.79	327.01	.00306	163.51	163.51	.99998	.51457	.21353	.21354	.99999
5.80	**330.30**	**.00303**	**165.15**	**165.15**	**.99998**	**2.51891**	**2.21787**	**2.21788**	**$\bar{1}$.99999**
5.81	333.62	.00300	166.81	166.81	.99998	.52325	.22222	.22222	.99999
5.82	336.97	.00297	168.48	168.49	.99998	.52759	.22656	.22657	.99999
5.83	340.36	.00294	170.18	170.18	.99998	.53194	.23090	.23091	.99999
5.84	343.78	.00291	171.89	171.89	.99998	.53628	.23525	.23525	.99999
5.85	347.23	.00288	173.62	173.62	.99998	.54062	.23959	.23960	.99999
5.86	350.72	.00285	175.36	175.36	.99998	.54497	.24393	.24394	.99999
5.87	354.25	.00282	177.12	177.13	.99998	.54931	.24828	.24828	.99999
5.88	357.81	.00279	178.90	178.91	.99998	.55365	.25262	.25262	.99999
5.89	361.41	.00277	180.70	180.70	.99998	.55799	.25696	.25697	.99999
5.90	**365.04**	**.00274**	**182.52**	**182.52**	**.99998**	**2.56234**	**2.26130**	**2.26131**	**$\bar{1}$.99999**
5.91	368.71	.00271	184.35	184.35	.99999	.56668	.26565	.26565	.99999
5.92	372.41	.00269	186.20	186.21	.99999	.57102	.26999	.27000	.99999
5.93	376.15	.00266	188.08	188.08	.99999	.57537	.27433	.27434	.99999
5.94	379.93	.00263	189.97	189.97	.99999	.57971	.27868	.27868	.99999
5.95	383.75	.00261	191.88	191.88	.99999	.58405	.28302	.28303	.99999
5.96	387.61	.00258	193.80	193.81	.99999	.58840	.28736	.28737	.99999
5.97	391.51	.00255	195.75	195.75	.99999	.59274	.29171	.29171	.99999
5.98	395.44	.00253	197.72	197.72	.99999	.59708	.29605	.29605	.99999
5.99	399.41	.00250	199.71	199.71	.99999	.60142	.30039	.30040	.99999
6.00	**403.43**	**.00248**	**201.71**	**201.72**	**.99999**	**2.60577**	**2.30473**	**2.30474**	**.99999**

14. BESSEL FUNCTIONS

The chart below shows approximate values for some representative Bessel functions of the first kind. Values for higher-order Bessel functions can be computed by successive application of the recurrence formula

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x)$$

starting with the values of $J_0(x)$ and $J_1(x)$.

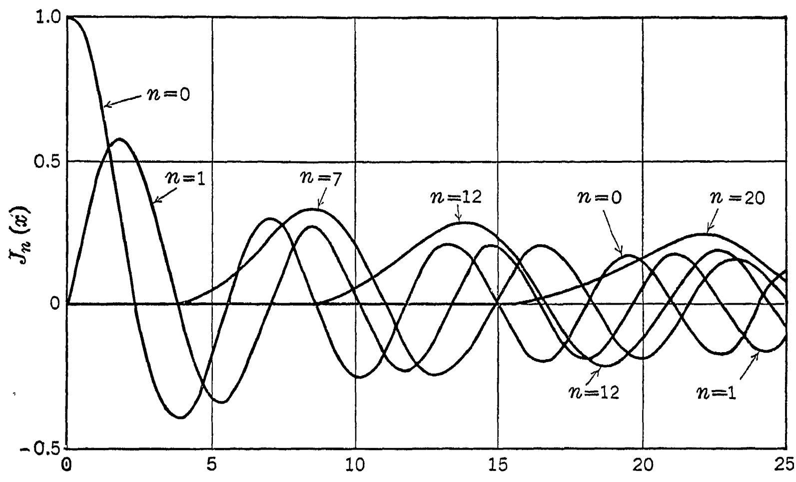


FIG. 1. Chart of Bessel Functions.

15. TRANSMISSION UNIT AND POWER REFERENCE LEVELS.—DECIBELS

Power losses occur in all parts of an electric circuit; in many circuits, which are built up of a number of components, the easiest method of predicting overall efficiencies is to determine individual efficiencies and combine them. When amplifiers are used, power gains exist (*as far as the alternating signal current is concerned*) and must be considered. For, although the useful power is less than the total power input, the output signal power may be greater than the input signal power, so that, using the conventional definition of efficiency, values greater than 100 per cent can be obtained.

THE DECIBEL AND THE NEPER. In order to avoid the multiplication of the individual efficiencies recourse has been had to logarithms of the efficiencies, giving measures of efficiency which can be added and subtracted directly.

Many units have been proposed and several have at various times been adopted in different localities. At the present time two such units are in general use in Europe and this country, one based on the napierian system of logarithms, the other based on the decimal system of logarithms. The International Advisory Committee on Long Distance Telephony of Europe has recommended that both these units be standardized and defined as follows:

The **unit of transmission** *expresses the ratio of apparent or real power in transmission systems. In practice, the number of units of transmission in a given case is expressed in terms of a logarithm.*

In the case of two powers P_1 and P_2 the number of units is:

$$\textit{in the napierian system,}\quad {}^1/_2 \log_\epsilon \frac{P_1}{P_2}$$

$$\textit{in the decimal system,}\quad \log_{10} \frac{P_1}{P_2}$$

The napierian unit is called the **neper.** The decimal unit is called the **bel.** A decimal submultiple of these units may be used, as **decineper** and **decibel.**

The unit generally used in this country is the decibel, which is exactly equivalent to and was first standardized as the "transmission unit"; it is also exactly equivalent to the "sensation unit" used in acoustic work. The decibel is abbreviated db, the transmission unit TU, and the sensation unit SU. The neper is called the βl unit. The comparative sizes of the two units is given by the fact that

$$\log_\epsilon \frac{P_1}{P_2} = 2.3026 \log_{10} \frac{P_1}{P_2},$$

so that 0.8686 neper is equal to 1 bel. Also 1 neper = 11.51 decibels.

From the above definition the number of decibels which expresses the ratio between any two powers is

$$N = 10 \log_{10} \frac{P_1}{P_2}$$

The quantity N is called the *transmission equivalent* of the element considered. It may be readily evaluated for a particular ratio by multiplying the common logarithm of this ratio by 10. If P_1 represents the delivered power and P_2 the input power, N will be negative for power losses and positive for power gains, since the logarithms of numbers less than unity are negative.

The use of the decibel may be seen from the following. If two circuit elements, with a ratio of power output to power input of $\frac{P_1}{P_2}$ and $\frac{P_3}{P_4}$, respectively, are connected in series, the power ratio of the combination is

$$\frac{P \text{ out}}{P \text{ in}} = \frac{P_1}{P_2} \times \frac{P_3}{P_4} = 10^{N_1(0.1)} \times 10^{N_2(0.1)} = 10^{(N_1+N_2)0.1}$$

where N_1 and N_2 are the transmission equivalents of the first and second elements, respectively. Taking the logarithms of both sides and multiplying through by 10,

$$10 \log \frac{P \text{ out}}{P \text{ in}} = N_1 + N_2 = N_T$$

It is thus seen that any number of transmission equivalents can be added (losses with their associated minus sign) to obtain the transmission equivalent of a complete circuit.

In making measurements of circuit efficiency the current ratio, or the voltage ratio, is usually more readily obtainable than the power ratio. Either of these ratios may be used to specify the efficiency of the circuit *when conditions are such that it is the square root of the power ratio.*

In this case

$$\frac{I_1^2}{I_2^2} = \frac{P_1}{P_2} = 10^{N(0.1)}$$

and taking the square root of both sides

$$\frac{I_1}{I_2} = 10^{N(0.05)}$$

By the method used above in the case of power ratios, the transmission equivalent is

$$N = 20 \log_{10} \frac{I_1}{I_2} = 20 \log_{10} \frac{E_1}{E_2}.$$

It must be remembered that this is true only when the current or voltage ratio is the square root of the power ratio, the simplest case being where the currents through, or voltages across, equal impedances are measured.

LOGARITHMIC VOLTAGE RATIO. In measuring electron-tube amplifiers, it is frequently useful to measure the voltage gain of each stage and compare the sum with the overall gain of the complete amplifier. Such measurements are in many cases conveniently made in terms of voltage and are made only with great difficulty in terms of power. The habit has grown up of using the advantages of logarithmic addition and calibrating amplifiers in terms of comparison voltages without any regard for the impedance relations. Furthermore in some instances the overall sensitivity of amplifiers and complete radio sets has frequently been expressed in terms of "decibels below 1 volt" with no thought of impedance. According to the above discussion this, of course, is a misuse of the term. Since it is so convenient, however, it is a practice which is likely to continue. Confusion can be avoided by the association of a new term to this measurement. It has been suggested that the abbreviation **dbv** be used to indicate that the logarithmic ratios are in terms of volts and not in terms of power. It has furthermore been suggested that the dbv also carry the implication where appropriate that it is below the level of 1 volt. A

Decibels Versus Power, Voltage, and Current Ratios

db	Current and Voltage Ratio		Power Ratio		db	Current and Voltage Ratio		Power Ratio	
	Gain	Loss	Gain	Loss		Gain	Loss	Gain	Loss
0.1	1.012	0.9886	1.023	0.9772	5.6	1.905	0.5248	3.631	0.2754
0.2	1.023	.9772	1.047	.9550	5.7	1.928	.5188	3.715	.2692
0.3	1.035	.9661	1.072	.9333	5.8	1.950	.5129	3.802	.2630
0.4	1.047	.9550	1.097	.9120	5.9	1.973	.5070	3.891	.2570
0.5	1.059	.9441	1.122	.8913	6.0	1.995	.5012	3.981	.2512
0.6	1.072	.9333	1.148	.8710	6.1	2.018	.4958	4.074	.2455
0.7	1.084	.9226	1.175	.8511	6.2	2.042	.4898	4.169	.2399
0.8	1.097	.9120	1.202	.8318	6.3	2.065	.4842	4.266	.2344
0.9	1.109	.9016	1.230	.8128	6.4	2.089	.4786	4.365	.2291
1.0	1.122	.8913	1.259	.7943	6.5	2.114	.4732	4.467	.2239
1.1	1.135	.8811	1.288	.7763	6.6	2.138	.4677	4.571	.2188
1.2	1.148	.8710	1.318	.7586	6.7	2.163	.4624	4.677	.2138
1.3	1.162	.8610	1.349	.7413	6.8	2.188	.4571	4.786	.2089
1.4	1.175	.8511	1.380	.7244	6.9	2.213	.4519	4.898	.2042
1.5	1.189	.8414	1.413	.7080	7.0	2.239	.4467	5.012	.1995
1.6	1.202	.8318	1.445	.6918	7.1	2.265	.4416	5.129	.1950
1.7	1.216	.8222	1.479	.6761	7.2	2.291	.4365	5.248	.1906
1.8	1.230	.8128	1.514	.6607	7.3	2.317	.4315	5.370	.1862
1.9	1.245	.8035	1.549	.6457	7.4	2.344	.4266	5.495	.1820
2.0	1.259	.7943	1.585	.6310	7.5	2.371	.4217	5.623	.1778
2.1	1.274	.7852	1.622	.6166	7.6	2.399	.4169	5.754	.1738
2.2	1.288	.7763	1.660	.6026	7.7	2.427	.4121	5.888	.1698
2.3	1.303	.7674	1.698	.5888	7.8	2.455	.4074	6.026	.1660
2.4	1.318	.7586	1.738	.5754	7.9	2.483	.4027	6.166	.1622
2.5	1.334	.7499	1.778	.5623	8.0	2.512	.3981	6.310	.1585
2.6	1.349	.7413	1.820	.5495	8.1	2.541	.3936	6.457	.1549
2.7	1.365	.7328	1.862	.5370	8.2	2.570	.3891	6.607	.1514
2.8	1.380	.7244	1.905	.5248	8.3	2.600	.3846	6.761	.1479
2.9	1.396	.7161	1.950	.5129	8.4	2.630	.3802	6.918	.1445
3.0	1.413	.7080	1.995	.5012	8.5	2.661	.3758	7.079	.1413
3.1	1.429	.6998	2.042	.4898	8.6	2.692	.3715	7.244	.1380
3.2	1.445	.6918	2.089	.4786	8.7	2.723	.3673	7.413	.1349
3.3	1.462	.6839	2.138	.4677	8.8	2.754	.3631	7.586	.1318
3.4	1.479	.6761	2.188	.4571	8.9	2.786	.3589	7.762	.1288
3.5	1.496	.6683	2.239	.4467	9.0	2.818	.3548	7.943	.1259
3.6	1.514	.6607	2.291	.4365	9.1	2.851	.3508	8.128	.1230
3.7	1.531	.6531	2.344	.4266	9.2	2.884	.3467	8.318	.1202
3.8	1.549	.6457	2.399	.4169	9.3	2.917	.3428	8.511	.1175
3.9	1.567	.6383	2.455	.4074	9.4	2.951	.3389	8.710	.1148
4.0	1.585	.6310	2.512	.3981	9.5	2.985	.3350	8.913	.1122
4.1	1.603	.6237	2.570	.3891	9.6	3.020	.3311	9.120	.1097
4.2	1.622	.6166	2.630	.3802	9.7	3.055	.3273	9.333	.1072
4.3	1.641	.6095	2.692	.3715	9.8	3.090	.3236	9.550	.1047
4.4	1.660	.6026	2.754	.3631	9.9	3.126	.3199	9.772	.1023
4.5	1.679	.5957	2.818	.3548	10.0	3.162	.3162	10.000	.1000
4.6	1.698	.5888	2.884	.3467	10.1	3.199	.3126	10.23	.0977
4.7	1.718	.5821	2.951	.3389	10.2	3.236	.3090	10.47	.0955
4.8	1.738	.5754	3.020	.3311	10.3	3.273	.3055	10.72	.0933
4.9	1.758	.5689	3.090	.3236	10.4	3.311	.3020	10.97	.0912
5.0	1.778	.5623	3.162	.3162	10.5	3.350	.2985	11.22	.0891
5.1	1.799	.5559	3.236	.3090	10.6	3.388	.2951	11.48	.0871
5.2	1.820	.5495	3.311	.3020	10.7	3.428	.2917	11.75	.0851
5.3	1.841	.5433	3.388	.2951	10.8	3.467	.2884	12.02	.0832
5.4	1.862	.5370	3.467	.2884	10.9	3.508	.2851	12.30	.0813
5.5	1.884	.5309	3.548	.2818	11.0	3.548	.2818	12.59	.0794

Decibels Versus Power, Voltage, and Current Ratios—*Continued*

db	Current and Voltage Ratio		Power Ratio		db	Current and Voltage Ratio		Power Ratio	
	Gain	Loss	Gain	Loss		Gain	Loss	Gain	Loss
11.1	3.589	0.2786	12.88	0.0776	16.1	6.383	0.1566	40.74	0.0245
11.2	3.631	.2754	13.18	.0759	16.2	6.457	.1549	41.69	.0239
11.3	3.673	.2723	13.49	.0741	16.3	6.531	.1531	42.66	.0234
11.4	3.715	.2692	13.81	.0724	16.4	6.607	.1514	43.65	.0229
11.5	3.758	.2661	14.13	.0708	16.5	6.683	.1496	44.67	.0224
11.6	3.802	.2630	14.45	.0692	16.6	6.761	.1479	45.71	.0219
11.7	3.846	.2600	14.79	.0676	16.7	6.839	.1462	46.77	.0214
11.8	3.891	.2570	15.14	.0661	16.8	6.918	.1445	47.86	.0209
11.9	3.936	.2541	15.49	.0646	16.9	6.998	.1429	48.98	.0204
12.0	3.981	.2512	15.85	.0631	17.0	7.079	.1413	50.12	.0200
12.1	4.027	.2483	16.22	.0617	17.1	7.161	.1396	51.29	.0195
12.2	4.074	.2455	16.60	.0603	17.2	7.244	.1380	52.43	.0191
12.3	4.121	.2427	16.98	.0589	17.3	7.328	.1365	53.70	.0186
12.4	4.169	.2399	17.38	.0575	17.4	7.413	.1349	54.96	.0182
12.5	4.217	.2371	17.78	.0562	17.5	7.499	.1334	56.23	.0178
12.6	4.266	.2344	18.20	.0550	17.6	7.586	.1318	57.54	.0174
12.7	4.315	.2317	18.62	.0537	17.7	7.674	.1303	58.88	.0170
12.8	4.365	.2291	19.05	.0525	17.8	7.762	.1288	60.26	.0166
12.9	4.416	.2265	19.50	.0513	17.9	7.852	.1273	61.66	.0162
13.0	4.467	.2239	19.95	.0501	18.0	7.943	.1259	63.10	.0158
13.1	4.519	.2213	20.42	.0490	18.1	8.035	.1245	64.57	.0155
13.2	4.571	.2188	20.89	.0479	18.2	8.128	.1230	66.07	.0151
13.3	4.624	.2163	21.38	.0468	18.3	8.222	.1216	67.61	.0148
13.4	4.677	.2138	21.88	.0457	18.4	8.318	.1202	69.18	.0145
13.5	4.732	.2113	22.39	.0447	18.5	8.414	.1189	70.80	.0141
13.6	4.786	.2089	22.91	.0437	18.6	8.511	.1175	72.44	.0138
13.7	4.842	.2065	23.44	.0427	18.7	8.610	.1161	74.13	.0135
13.8	4.898	.2042	23.99	.0417	18.8	8.710	.1148	75.86	.0132
13.9	4.955	.2018	24.55	.0407	18.9	8.811	.1135	77.63	.0129
14.0	5.012	.1995	25.12	.0398	19.0	8.913	.1122	79.43	.0126
14.1	5.070	.1972	25.70	.0389	19.1	9.016	.1109	81.28	.0123
14.2	5.129	.1950	26.30	.0380	19.2	9.120	.1097	83.18	.0120
14.3	5.188	.1928	26.92	.0372	19.3	9.226	.1084	85.11	.0117
14.4	5.248	.1906	27.54	.0363	19.4	9.333	.1072	87.10	.0115
14.5	5.309	.1884	28.18	.0355	19.5	9.441	.1059	89.13	.0112
14.6	5.370	.1862	28.84	.0347	19.6	9.550	.1047	91.20	.0110
14.7	5.433	.1841	29.51	.0339	19.7	9.661	.1035	93.33	.0107
14.8	5.495	.1820	30.20	.0331	19.8	9.772	.1023	95.50	.0105
14.9	5.559	.1799	30.90	.0324	19.9	9.886	.1012	97.72	.0102
15.0	5.623	.1778	31.62	.0316	20.0	10.000	.1000	100.0	.0100
15.1	5.689	.1758	32.36	.0309	30.0	31.62	.0316	1,000	.0010
15.2	5.754	.1738	33.11	.0302	40.0	100.0	.0100	10^4	10^{-4}
15.3	5.821	.1718	33.88	.0295	50.0	316.2	.0032	10^5	10^{-5}
15.4	5.888	.1698	34.67	.0288	60.0	1,000.0	.0010	10^6	10^{-6}
15.5	5.957	.1679	35.48	.0282	70.0	3,162.0	.0003	10^7	10^{-7}
15.6	6.026	.1660	36.31	.0275	80.0	10,000.0	.0001	10^8	10^{-8}
15.7	6.096	.1641	37.15	.0269	90.0	31,620.0	.00003	10^9	10^{-9}
15.8	6.166	.1622	38.02	.0263	100.0	100,000.0	.00001	10^{10}	10^{-10}
15.9	6.237	.1603	38.91	.0257					
16.0	6.310	.1585	39.81	.0251					

sensitivity of 100 μv could be expressed as -80 dbv, and a sensitivity of 10 μv as -100 dbv, with this system.

A table is appended giving values of transmission equivalents in terms of both power and current, or voltage, ratios in tenths of a decibel up to 20 db. For values above 20 db the tables may be used as described below.

Example. To find the current and power ratios for a loss of 57.6 db.

1. The power ratio of 50 db is 10^{-5} (this being the first power ratio which is an even submultiple of 10 and corresponds to less than 57.6 db).
2. The power ratio of 7.6 ($57.6 - 50 = 7.6$) db is 0.1738.
3. To add decibels, power ratios must be multiplied, hence:

$$\textit{Power ratio of } 57.6 \textit{ db} = 0.1738 \times 10^{-5}$$

4. The current ratio of 40 db is 0.01 (first current ratio which is an even submultiple of 10 and corresponds to less than 57.6 db).
5. The current ratio of 17.6 ($57.6 - 40 = 17.6$) db is 0.1318.
6. Multiplying these ratios:

$$\textit{Current ratio of } 57.6 \textit{ db} = 0.001318$$

POWER REFERENCE LEVELS. When the efficiency of a device or system is expressed in decibels there is in general no indication of the actual amount of power in the device. In comparing devices it is frequently desirable to know the actual overall efficiency. In such a case this can readily be expressed in decibels, 100 per cent efficiency being represented by zero decibels. In many cases, however, it is more desirable that the relative efficiencies at different frequencies be known; in some such cases it is also desirable that the normal power capacity of the device be specified in such form as to be readily comparable with similar devices.

For such a specification to be made when the ordinates of a characteristic are in decibels, it is only necessary to specify some arbitrary amount of power as corresponding to zero decibels; then every value of decibels represents a definite amount of power (or volume level). The amount of power chosen as the reference level is completely arbitrary; hence it has been customary to choose some average value of power as zero level, for a particular type of work.

Attempts at standardization have been made, with the result that the American Standards Association (and IRE) has recommended 1 milliwatt in 600 ohms in connection with a particular meter to measure levels in radio program transmission (see ASA "American Recommended Practice for Volume Measurements of Electrical Speech and Program Waves," Nov. 6, 1942) and has introduced the term vu (vee-you) to represent the number of decibels above or below this level. Also it has become customary to specify power in *dbm* which is used to mean decibels above or below 1 milliwatt.

However, some other groups are still using other levels. For instance, in sound-motion pictures the reference level is 6 milliwatts. Also certain radar engineers use decibels below or above 1 watt.

It will apparently require further action by the American Standards Association to bring order out of the present chaos. The desirability of such a standard is shown by the experience in the acoustic field where 10^{-16} watt per sq cm was universally adopted, so that measurements made anywhere are everywhere intelligible. Until the adoption of such a standard, great care must be exercised in comparing curves and statements of levels to insure that correction is made for differences in reference levels. Also a statement of the reference used should always be included as a part of any publication of results.

BIBLIOGRAPHY

Hilliard, J. K., Definition of Standard Reference Systems, *Electronics*, Vol. 3, 192 (November 1931).
Martin, W. H., Decibel—The Name for the Transmission Unit, *B.S.T.J.*, Vol. 8, 1 (January 1929).
Martin, W. H., and C. H. G. Gray, Master Reference System for Telephone Transmission, *B.S.T.J.*, Vol. 8, 536 (July 1929).

UNITS AND CONVERSION FACTORS

By J. G. Brainerd and Carl C. Chambers

16. SYSTEMS OF UNITS

The magnitude of a physical quantity has no tangible meaning except as the relative magnitude of that quantity as compared with some other quantity of the same nature. Thus, 50 ohms is a resistance having a magnitude 50 times the resistance of 1 ohm. Therefore, whenever it is necessary or desirable to talk about the magnitude of a physical quantity, it is necessary to have a basis for comparison. This basis for a quantity is called the *unit* of magnitude of that quantity. In order to communicate the idea of magnitude between different people, it is necessary that they at least know the relative magnitudes of their units. It is the purpose of this section to act as tool for the specification of the relative magnitudes of the more commonly used systems of units for physical quantities.

Because of the relations defining physical laws, there are relations between the magnitudes of physical quantities. It is desirable that these physical relations be expressed alike in the different systems of units. For instance, the relation mass × acceleration = force should be independent of the system of units. Therefore, unit mass times unit acceleration should equal unit force. This gives a relation among these three units.

Because of such physical relations, all the mechanical units can be derived from the units for three fundamental quantities. The three quantities ordinarily taken as fundamental are mass, length, and time. Thermal quantities are conveniently derived from these three quantities together with another fundamental quantity, temperature. Photometric quantities are derived from the three fundamental mechanical quantities together with luminous intensity as a fourth fundamental quantity.

Similarly, electrical and magnetic quantities are derived from the three fundamental mechanical quantities and one fundamental electrical or magnetic quantity.

Two systems of mechanical units are in use in English-speaking countries, the English and the metric systems. The metric system is used universally by physicists and to a great extent by engineers, although the English system is still very common in engineering. The English system uses the foot, the pound, and the second as the units for length, mass, and time, respectively. The metric system (as used in the current literature—see MKS system below) employs the meter, the kilogram, and the second as the units for length, mass, and time, respectively.

STANDARDS OF THE FUNDAMENTAL UNITS. The physical units upon which these fundamental units are based and the legalized standards of the United States and Great Britain are described below.

Standard of Length. The standard meter (100 cm) is the distance between two lines on a platinum-iridium bar carefully preserved at the Bureau of Weights and Measures, at Sèvres, France, when the bar is kept at a uniform temperature of 0 deg cent throughout. In the United States the yard (3 ft) was defined by Act of Congress, July 28, 1866, as

$$1 \text{ U. S. yard} = \frac{3600}{3937} \text{ meter}$$

and similarly the British imperial yard is defined by law as

$$1 \text{ British imperial yard} = \frac{3600}{3937.079} \text{ meter}$$

For engineering purposes the U. S. and British yards may be considered identical.

Standard of Mass and Force. The standard kilogram (1000 grams) is the mass of a cylinder of platinum preserved at the Bureau of Weights and Measures, at Sèvres, France. The U. S. pound avoirdupois is defined by law (Act of Congress, 1866) as $\frac{1}{2.2046}$ kg, but in 1893, the Superintendent of Weights and Measures, with the approval of the Secretary of the Treasury, declared the pound to be

$$1 \text{ U. S. lb} = \frac{1}{2.204622} \text{ kg}$$

The British imperial pound has the same value.

The same relations between the pound and kilogram hold whether these units be taken as units of mass or as units of force, the unit of force being defined in both cases as the pull of the earth on unit mass at 45 deg latitude and sea level.

Standard of Time. The standard second universally adopted is the $\frac{1}{86{,}400}$ part of a mean solar day. The solar day is the interval of time between two successive transits of the sun across a meridian of the earth at the point of observation; this interval varies in length at different times during the year, but the average length of the interval for one year is constant as far as can be determined by any known methods of observation.

Temperature Scales. Two units of temperature, or temperature scales, are commonly employed, viz., the centigrade and the fahrenheit units. The relation between these two units results solely from the manner in which they are defined. One degree centigrade $= {}^9/_5$ degree fahrenheit. Owing to the difference in the zeros of the two scales, a temperature of t_f degrees fahrenheit corresponds to a temperature of $t_c = {}^5/_9(t_f - 32)$ degrees centigrade, and vice versa, $t_f = {}^9/_5\,t_c + 32$ degrees fahrenheit.

Standard of Luminous Intensity. Before Jan. 1, 1948, the standard of luminous intensity was the mean intensity in the horizontal plane from a group of incandescent lamps maintained by the National Bureau of Standards (U. S.), in cooperation with similar custodians in France, Great Britain, and Germany. The International candle was a point source of light having an intensity of a definite fraction of this standard intensity.

The National Bureau of Standards, in pursuance of decisions of the International Committee on Weights and Measures, decided that, beginning Jan. 1, 1948, it would take as the primary standard for the system of photometric units a black-body radiator operated at the temperature of freezing platinum. The "candle," unit of intensity, is defined as one-sixtieth of the intensity of one square centimeter of such a radiator. Other units are derived from the candle, with the provision that when differences of color are involved the evaluation shall be made by means of standard spectral luminosity factors which have been adopted by the International Commission on Illumination and the International Committee on Weights and Measures.

ELECTRIC UNITS. Three systems of electric and magnetic units are in general use, viz., (1) the cgs electrostatic system, (2) the cgs electromagnetic system, and (3) the practical system. In the cgs electrostatic system the dielectric coefficient, ϵ, of air * at 0 deg cent and 760 mm mercury pressure is arbitrarily chosen as unity. In the cgs electromagnetic system the magnetic permeability of air under the same standard conditions is arbitrarily chosen as unity. In the practical system a concrete standard of the unit of resistance (called the ohm) and of the unit of current (ampere) is arbitrarily chosen (it was stated above that only one electric or magnetic unit need be chosen; the choice of two leads to inconsistencies; see below); the unit of resistance is closely equal to 10^9 times the unit of resistance in the cgs electromagnetic system and the unit current is approximately 0.1 that in the latter system. Occasionally other (special) systems are used, most of which are designed to get rid of a factor 4π which frequently appears in the usual systems. The most popular of these others is the Heaviside–Lorentz system in which the unit of electric charge is $1/\sqrt{4\pi}$ of the unit in the electrostatic system. (See MKS system.)

Use of the Prefixes "Stat" and "Ab." To designate the electric and magnetic units in the electrostatic and electromagnetic systems of units respectively, the prefixes "stat" and "ab" may be used with the name of the corresponding practical unit. For example, the cgs electrostatic unit of electric charge may be called the statcoulomb and the cgs electromagnetic unit of electric charge may be called the abcoulomb, etc.†

Relations among the Three Systems of Electrical Units. The fundamental relations, experimentally determined, between the cgs electrostatic and the cgs electromagnetic system is that 1 abfarad $= 8.9878 \times 10^{20}$ statfarads, which may be approximated for engineering purposes to

$$1 \text{ abfarad} = 9 \times 10^{20} \text{ statfarads}$$

which, as a consequence of the definition of the various terms, is equivalent to

$$1 \text{ abcoulomb} = 3 \times 10^{10} \text{ statcoulombs}$$

the erg being the unit of energy in both systems. Rigorously,

$$1 \text{ abcoulomb} = 2.9979 \times 10^{10} \text{ statcoulombs}$$

(See the article by Birge, *Rev. of Mod. Phys.*, Vol. 1, 1 [July 1929].)

---

* Rigorously, ϵ_0 of free or empty space is chosen unity; for air at 0 deg cent and 76 cm mercury pressure $\epsilon_0 = 1.000585$; see *International Critical Tables*, Vol. 6, 77, for the value of ϵ for air under various conditions.

† This abcoulomb, the unit of quantity of electricity in the electromagnetic system, should not be confused with an "absolute coulomb," which is a unit closely equal to the coulomb and is what the latter would be if 1 international or practical ohm equaled *exactly* 10^9 abohms and 1 ampere equaled *exactly* 0.1 abamp. For engineering purposes, the difference between an absolute coulomb and a coulomb is negligible.

The fundamental relations between the cgs electromagnetic system and the practical system are

$$1 \text{ abcoulomb} = 10 \text{ coulombs}$$

$$1 \text{ erg} = 10^{-7} \text{ watt-second or joule}$$

the erg being the unit of energy in the cgs electromagnetic system and the joule (or watt-second) that in the practical system.

Practical Electrical Units. The former (see below) legal units of electrical measure in the United States are given in an Act of Congress, July 12, 1894. Unfortunately, the units there defined are not consistent with one another; for example, the unit of power (watt) there given is not equal to the unit of power derived from the units of current (ampere) and voltage (volt) as defined in the Act. The practical units (the so-called international units) in use before Jan. 1, 1948, are based on the following two definitions:

The unit of resistance is the (international) ohm and is equal to the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice, 14.4521 grams in mass, of a constant cross-sectional area and 106.300 cm in length.

The unit of current is the (international) ampere and is equal to the unvarying electric current which, when passed through a solution of nitrate of silver in accordance with certain specifications, deposits silver at the rate of 0.00111800 gram per second.

The unit of electromotive force, the (international) volt, is derived from the above by Ohm's law. Other international units are derived from these.

The National Bureau of Standards, in agreement with decisions of the International Committee on Weights and Measures, decided to use as standard, beginning Jan. 1, 1948, the electrical units "derived from the fundamental mechanical units of length, mass, and time by use of accepted principles of electromagnetism, with the value of the permeability of space taken as unity in the centimeter-gram-second system or as 10^{-7} in the corresponding meter-kilogram-second system." The reference is to the unrationalized MKS system; in the rationalized MKS system, the permeability of space is $4\pi \times 10^{-7}$ henry per meter.

In explanation of the legal status of the new standard, the Bureau states, "When the electrical units were defined by law (Public Law No. 105, 53rd Congress) in 1894 it was supposed that the international units were practically identical with the corresponding multiples of the centimeter-gram-second electromagnetic system. Alternative definitions were given for most of the units, and those definitions which appear to be legally controlling were taken partly from one system and partly from the other. The joule and the watt, for example, are clearly defined as multiples of the cgs units. In brief, the absolute units have as good a legal basis under the terms of that act as do the present international units. New legislation is being proposed to remove the ambiguities of the old act, but there should be no objection on legal grounds to the general adoption of the absolute units even in advance of Congressional action."

Using "international" to refer to the previous standard, and "absolute" to refer to the new, the relations accepted by the International Committee on Weights and Measures at its meeting in Paris in October, 1946, are as follows:

$$1 \text{ mean international ohm} = 1.00049 \text{ absolute ohms}$$

$$1 \text{ mean international volt} = 1.00034 \text{ absolute volts}$$

The mean international units to which the above equations refer are the averages of units as maintained in the national laboratories of the six countries (France, Germany, Great Britain, Japan, U.S.S.R., and the United States) which took part in this work before the war. The units maintained by the National Bureau of Standards differ from these average units by a few parts in a million, so that the conversion factors for adjusting values of standards in this country will be as follows:

$$1 \text{ international ohm (U. S.)} = 1.000495 \text{ absolute ohms}$$

$$1 \text{ international volt (U. S.)} = 1.00033 \text{ absolute volts.}$$

Other electrical units will be changed by amounts shown in the following table:

$$1 \text{ international ampere} = 0.999835 \text{ absolute ampere}$$

$$1 \text{ international coulomb} = 0.999835 \text{ absolute coulomb}$$

$$1 \text{ international henry} = 1.000495 \text{ absolute henrys}$$

$$1 \text{ international farad} = 0.999505 \text{ absolute farad}$$

$$1 \text{ international watt} = 1.000165 \text{ absolute watts}$$

$$1 \text{ international joule} = 1.000165 \text{ absolute joules}$$

The Act of 1894 defined the international ohm as previously stated, but defined the ampere as 0.1 abampere. These units give rise to the so-called "semiabsolute" system, which is seldom used.

THE MKS SYSTEM OF UNITS. In 1904 Giorgi proposed a system of units in which the fundamental units were the meter, the kilogram, the second, and the ohm. Using this system of fundamental units, the permeability of free space is $\mu_0 = 4\pi \times 10^{-7}$ henry per meter, and the equations of electricity and magnetism, using the practical units, become equations without factors such as 10^8, etc. Such a system is similar to the so-called absolute systems such as the cgs electromagnetic and the cgs electrostatic systems. It follows from the theory of radiation of electromagnetic waves that the dielectric coefficient $\epsilon_0 = \dfrac{1}{\mu_0 c^2}$, where c is the ratio of electromagnetic to electrostatic units, which can be taken as the velocity of light in free space.

The International Committee of Weights and Measures, at its meeting in October 1946, decided that the actual substitution of this absolute system of electrical units for the international system should take place on January 1, 1948.

The units are then defined by a set of definitions such as follows:

(a) **Ampere.** The ampere is the constant current which, maintained in two parallel rectilinear conductors of infinite length separated by a distance of 1 meter, produces between these conductors a force equal to 2×10^{-7} mks (meter-kilogram-second) units of force per meter of length.

(b) **Volt.** The volt is the difference of electrical potential between two points of a conductor carrying a constant current of 1 ampere when the power dissipated between these points is equal to 1 mks unit of power (watt).

(c) **Coulomb.** The coulomb is the quantity of electricity transported each second by a current of 1 ampere.

(d) **Ohm.** The ohm is the electrical resistance between two points of a conductor when a constant difference of potential of 1 volt, applied between these points, produces in the conductor a current of 1 ampere, the conductor not being the seat of an electromotive force.

(e) **Weber.** The weber is the magnetic flux which, traversing a circuit of a single turn, would produce an electromotive force of 1 volt, if brought to zero in 1 second with uniform diminution.

(f) **Henry.** The henry is the inductance of a closed circuit in which an electromotive force of 1 volt is produced when the electric current traversing the circuit varies uniformly at the rate of 1 ampere per second.

(g) **Farad.** The farad is the electrical capacitance of a capacitor between the plates of which appears an electrical difference of potential of 1 volt, when charged with 1 coulomb of electric charge.

The original Giorgi MKS system chose the ohm as the fourth fundamental unit. This choice has not been confirmed. The electrical fundamental unit could be almost any of the electrical units. No particular unit has as yet been chosen as fundamental. The preferences seem to be divided between the ampere, the ohm, the permeability, and the coulomb.

The original Giorgi MKS system chose $\mu_0 = 4\pi 10^{-7}$ henry per meter, the 4π factor causing the electromagnetic formulas expressing rectilinear symmetry, such as the Maxwell equations, to be free of the factor 4π, and the electromagnetic formulas expressing circular symmetry, such as Coulomb's law, to contain the factor 4π. Such a system is called a rationalized system as contrasted with a non-rationalized system, examples of which are the electromagnetic and the electrostatic cgs systems. The non-rationalized MKS system corresponding to the original Giorgi system is defined by the choice of $\mu_0 = 10^{-7}$. This changes the values of some of the units as shown in the table below.

Rationalized MKS Units and Corresponding CGS Electromagnetic Units

Multiply mks units by F to obtain cgs units

Quantity	*Symbol*	MKS *Unit*	CGS *Unit*	F
Mechanical				
Length	L	m	cm	10^2
Mass	M	kg	g	10^3
Time	T	sec	sec	1
Area	S	sq m	sq cm	10^4
Volume	V	cu m (stere)	cu cm	10^6
Frequency	f	cycle per sec (hertz)	cycle per sec	1
Density	d	kg per cu m	g per cu cm	10^{-3}
Velocity	v	m per sec	cm per sec	10^2
Acceleration	a	m per sec per sec	cm per sec per sec	10^2
Force	F	newton (j per m)	dyne	10^5
Pressure	p	newton per sq m	dyne per sq cm	10
Angle	α, β	radian	radian	1
Angular velocity	ω	radian per sec	radian per sec	1
Torque	τ	j per radian	dyne cm	10^7
Moment of inertia	J	kg-sq m	g-sq cm	10^7
Energetics				
Work or energy	W	j	erg	10^7
Volume energy or energy density	w	j per cu m	erg per cu cm	10
Active power	P	w	erg per sec	10^7
Reactive power	Q	var	erg per sec	10^7
Thermal				
Quantity of heat	Q	kg cal	g cal	10^3
Temperature	θ	C or K	C or K	1
Luminous				
Intensity	I	candle	candle	1
Luminous flux	ψ	l	l	1
Illumination	E	lux	phot	10^{-4}
Brightness	b	candle per sq m	stilb	10^{-4}
Electrical				
Electromotive force	E	volt	abvolt	10^8
Potential gradient or electric field intensity	E	volt per m	abvolt per cm	10^6
Resistance	R	ohm	abohm	10^9
Resistivity	ρ	ohm-m	abohm-cm	10^{11}
Conductance	G	siemens, mho	abmho	10^{-9}
Conductivity	γ	mho per m	abmho per cm	10^{-11}
Quantity or displacement	Q	coulomb	abcoulomb	10^{-1}
Current	I	amp	abamp	10^{-1}
Electric flux	Ψ	coulomb	abcoulomb	10^{-1}
Flux density	D	coulomb per sq m	abcoulomb per sq cm	10^{-5}
Current density	i	ampere per sq m	abampere per sq cm	10^{-5}
Capacitance	C	farad	abfarad	10^{-9}
Specific inductive capacity	ϵ/ϵ_0	numeric	numeric	1
Dielectric coefficient for free space or space capacitivity	ϵ_0	$10^7/4\pi c^2 = 8.854 \times 10^{-12}$	$\frac{1}{c^2} = 1.113 \times 10^{-21}$	
Magnetic				
Magnetomotive force	$\mathfrak{F}$	amp-turn	gilbert	$4\pi 10^{-1}$
Magnetizing force or magnetic field intensity	H	amp-turn per m	oersted	$4\pi 10^{-3}$
Space permeability	μ_0	$4\pi 10^{-7} = 1.257 \times 10^{-6}$	1	
Relative permeability	μ/μ_0	numeric	numeric	1
Magnetic flux	Φ	weber	maxwell	10^8
Flux density	B	weber per sq m	gauss	10^4
Reluctance	$\mathcal{R}$	amp-turn per weber		$4\pi 10^{-9}$
Permeance	$\mathcal{P}$	weber per amp-turn		$10^9/4\pi$
Inductance	L	henry	abhenry	10^9
Pole strength	m	weber	maxwell/4π	$10^8/4\pi$
Magnetization	$\mathfrak{J}$	weber per sq m		$10^4/4\pi$
Magnetic moment	$\mathfrak{M}$	weber-m	maxwell-cm/4π	$10^{10}/4\pi$

Rationalized MKS Units and Corresponding Non-rationalized Units

Multiply non-rationalized mks units by F to obtain rationalized mks units

Quantity	*Symbol*	*Name of Rationalized* MKS *Units*	F
Electrical			
Electric flux	Ψ	coulomb	4π
Flux density	D	coulomb per sq m	4π
Space capacitivity	ϵ_0	farad per m	4π
Magnetic			
Magnetomotive force	M or $\mathfrak{F}$	amp-turn	$1/4\pi$
Magnetizing force	H	amp-turn per m	$1/4\pi$
Space permeability	μ_0	henry per m	4π
Permeance	$\mathcal{P}$	weber per amp-turn	4π
Reluctance	$\mathcal{R}$	amp-turn per weber	$1/4\pi$
Pole strength	m	weber	4π
Magnetic moment	$\mathfrak{M}$	weber-m	4π
Flux density	B	weber per sq m	4π

17. CONVERSION TABLES

Table 1. Length [L]

Multiply Number of → by to Obtain ↓	Centimeters	Feet	Inches	Kilometers	Nautical miles	Meters	Mils	Miles	Millimeters	Yards
Centimeters	1	30.48	2.540	10^5	1.853×10^5	100	2.540×10^{-3}	1.609×10^5	0.1	91.44
Feet	3.281×10^{-2}	1	8.333×10^{-2}	3281	6080.27	3.281	8.333×10^{-5}	5280	3.281×10^{-3}	3
Inches	0.3937	12	1	3.937×10^4	7.296×10^4	39.37	0.001	6.336×10^4	3.937×10^{-2}	36
Kilometers	10^{-5}	3.048×10^{-4}	2.540×10^{-5}	1	1.853	0.001	2.540×10^{-8}	1.609	10^{-6}	9.144×10^{-4}
Nautical miles		1.645×10^{-4}		0.5396	1	5.396×10^{-4}		0.8684		4.934×10^{-4}
Meters	0.01	0.3048	2.540×10^{-2}	1000	1853	1		1609	0.001	0.9144
Mils	393.7	1.2×10^4	1000	3.937×10^7		3.937×10^4	1		39.37	3.6×10^4
Miles	6.214×10^{-6}	1.894×10^{-4}	1.578×10^{-5}	0.6214	1.1516	6.214×10^{-4}		1	6.214×10^{-7}	5.682×10^{-4}
Millimeters	10	304.8	25.40	10^6		1000	2.540×10^{-2}		1	914.4
Yards	1.094×10^{-2}	0.3333	2.778×10^{-2}	1094	2027	1.094	2.778×10^{-5}	1760	1.094×10^{-3}	1

Metric Multiples

10^6 microns = 10^3 millimeters = 10^2 centimeters = 10 decimeters = 1 meter = 10^{-1} dekameter = 10^{-2} hectometer = 10^{-3} kilometer = 10^{-4} myriameter = 10^{-6} megameter = 10^{10} Angstrom Units.

Land Measure

7.92 inches = 1 link
25 links = 1 rod = 16.5 feet = 5.5 yards (1 rod = 1 pole = 1 perch)
4 rods = 1 chain (Gunther's) = 66 feet = 22 yards = 100 links
10 chains = 1 furlong = 660 feet = 220 yards = 1000 links = 40 rods
8 furlongs = 1 mile = 5280 feet = 1760 yards = 8000 links = 320 rods = 80 chains

Ropes and Cables

2 yards = 1 fathom   120 fathoms = 1 cable's length

Nautical Measure

6080.27 feet = 1 knot = 1 nautical mile = 1.15156 statute miles
3 nautical miles = 1 league (U. S.)   3 statute miles = 1 league (Gr. Britain)

(NOTE.—A knot, or nautical mile, is the length of a minute of longitude of the earth at the equator at sea level. The British Admiralty uses the round figure of 6080 feet. The word "knot" is frequently used also to denote "nautical miles per hour.")

Miscellaneous

3 inches = 1 palm   9 inches = 1 span
4 inches = 1 hand $2\,^1/_2$ feet = 1 military pace

Table 2. Area $[L^2]$

Multiply Number of → by to Obtain ↓	Acres	Circular mils	Square centimeters	Square feet	Square inches	Square kilometers	Square meters	Square miles	Square millimeters	Square yards
Acres	1			2.296×10^{-5}		247.1	$2\ 471 \times 10^{-4}$	640		2.066×10^{-4}
Circular mils		1	1.973×10^{5}	1.833×10^{8}	1.273×10^{6}		1.973×10^{9}		1973	
Square centimeters		5.067×10^{-6}	1	929.0	6.452	10^{10}	10^{4}	2.590×10^{10}	0.01	8361
Square feet	4.356×10^{4}		1.076×10^{-3}	1	6.944×10^{-3}	1.076×10^{7}	10.76	2.788×10^{7}	1.076×10^{-5}	9
Square inches	6,272,640	7.854×10^{-7}	0.1550	144	1	1.550×10^{9}	1550	4.015×10^{9}	1.550×10^{-3}	1296
Square kilometers	4.047×10^{-3}		10^{-10}	9.290×10^{-8}	6.452×10^{-10}	1	10^{-6}	2.590	10^{-12}	8.361×10^{-7}
Square meters	4047		0.0001	9.290×10^{-2}	6.452×10^{-4}	10^{6}	1	2.590×10^{6}	10^{-6}	0.8361
Square miles	1.562×10^{-3}		3.861×10^{-11}	3.587×10^{-8}		0.3861	$3\ 861 \times 10^{-7}$	1	3.861×10^{-13}	3.228×10^{-7}
Square millimeters		5.067×10^{-4}	100	9.290×10^{4}	645.2	10^{12}	10^{6}		1	8.361×10^{5}
Square yards	4840		1.196×10^{-4}	0.1111	7.716×10^{-4}	1.196×10^{6}	1.196	3.098×10^{6}	1.196×10^{-6}	1

Land Measure

$30\ ^1/_4$ square yards = 1 square rod = $272\ ^1/_4$ square feet
16 square rods = 1 square chain = 484 square yards = 4356 square feet
$2\ ^1/_2$ square chains = 1 rood = 40 square rods = 1210 square yards
4 roods = 1 acre = 10 square chains = 160 square rods
640 acres = 1 square mile = 2560 roods = 102,400 square rods
1 section of land = 1 square mile; 1 quarter section = 160 acres

Architect's Measure

100 square feet = 1 square

Circular Inch and Circular Mil

A circular inch is the area of a circle 1 inch in diameter = 0.7854 square inch
1 square inch = 1.2732 circular inches
A circular mil is the area of a circle 1 mil (or 0.001 inch) in diameter = 0.7854 square mil
1 square mil = 1.2732 circular mils
1 circular inch = 10^6 circular mils = 0.7854×10^6 square mils
1 square inch = 1.2732×10^6 circular mils = 10^6 square mils

Metric Multiples

1 square meter = 1 centiare = 10^{-2} are = 10^{-4} hectare
= 10^{-6} square kilometer = 10^{-8} square myriameter

Table 3. Volume [L^3]

Multiply Number of → by to Obtain ↓	Bushels (dry)	Cubic centimeters	Cubic feet	Cubic inches	Cubic meters (steres)	Cubic yards	Gallons (liquid)	Liters	Pints (liquid)	Quarts (liquid)
Bushels (dry)	1		0.8036	4.651×10^{-4}	28.38			2.838×10^{-2}		
Cubic centimeters	3.524×10^{4}	1	2.832×10^{4}	16.39	10^{6}	7.646×10^{5}	3785	1000	473.2	946.4
Cubic feet	1.2445	3.531×10^{-5}	1	5.787×10^{-4}	35.31	27	0.1337	3.531×10^{-2}	1.671×10^{-2}	3.342×10^{-2}
Cubic inches	2150.4	6.102×10^{-2}	1728	1	6.102×10^{4}	46,656	231	61.02	28.87	57.75
Cubic meters (steres)	3.524×10^{-2}	10^{-6}	2.832×10^{-2}	1.639×10^{-5}	1	0.7646	3.785×10^{-3}	0.001	4.732×10^{-4}	9.464×10^{-4}
Cubic yards		1.308×10^{-6}	3.704×10^{-2}	2.143×10^{-5}	1.308	1	4.951×10^{-3}	1.308×10^{-3}	6.189×10^{-4}	1.238×10^{-3}
Gallons (liquid)		2.642×10^{-4}	7.481	4.329×10^{-3}	264.2	202.0	1	0.2642	0.125	0.25
Liters	35.24	0.001	28.32	1.639×10^{-2}	1000	764.6	3.785	1	0.4732	0.9464
Pints (liquid)		2.113×10^{-3}	59.84	3.463×10^{-2}	2113	1616	8	2.113	1	2
Quarts (liquid)......		1.057×10^{-3}	29.92	1.732×10^{-2}	1057	807.9	4	1.057	0.5	1

Acre-feet: multiply number of acre-feet by 4.356×10^4 to obtain number of cubic feet; multiply by 3.259×10^5 to obtain number of gallons.

Metric Multiples

10 milliliters	= 1 centiliter	= 0.338 fluid ounce
10 centiliters	= 1 deciliter	= 0.845 liquid gill
10 deciliters	= 1 liter	= 1.0567 liquid quarts
10 liters	= 1 dekaliter	= 2.6417 liquid gallons
10 dekaliters	= 1 hectoliter	= 2.8375 U. S. bushels
10 hectoliters	= 1 kiloliter (or stere)	= 28.375 U. S. bushels

Cubic Measure

1 cord of wood = a pile cut 4 feet long, piled 4 feet high and 8 feet on the ground = 128 cubic feet

1 perch of stone = a quantity 1 1/2 feet thick, 1 foot high and 16 1/2 feet long = 24 3/4 cubic feet

(NOTE.—A perch of stone is, however, often computed differently in different localities; thus, in most if not all of the States and Territories west of the Mississippi, stonemasons figure rubble by the perch of 16 1/2 cubic feet. In Philadelphia, 22 cubic feet are called a perch. In Chicago, stone is measured by the cord of 100 cubic feet. Check should be made against local practice.)

Board Measure

In board measure, boards are assumed to be one inch in thickness. Therefore, feet board measure of a stick of square timber = length in feet × breadth in feet × thickness in inches.

Shipping Measure

For register tonnage or measurement of the entire internal capacity of a vessel, it is arbitrarily assumed, to facilitate computation, that:

100 cubic feet = 1 register ton

For the measurement of cargo:

40 cubic feet = 1 U. S. shipping ton = 32.143 U. S. bushels

42 cubic feet = 1 British shipping ton = 32.703 Imperial bushels

Dry Measure

One U. S. Winchester bushel contains 1.2445 cubic feet or 2150.42 cubic inches. It holds 77.601 pounds distilled water at 62 deg fahr.

(NOTE.—The above is a *struck* bushel. A *heaped* bushel in general equals 1 1/4 struck bushels, although for apples and pears it contains 1.2731 struck bushels = 2737.72 cubic inches.)

One U. S. gallon (dry measure) = 1/8 bushel and contains 268.8 cubic inches.

(NOTE.—This is not a legal U. S. *dry measure* and therefore is given for comparison only.)

One British Imperial bushel contains 1.2843 cubic feet or 2219.36 cubic inches. It holds 80 pounds distilled water at 62 deg fahr.

One British Imperial gallon = 1/8 Imperial bushel and contains 277.42 cubic inches.

1 Winchester bushel = 0.9694 Imperial bushel

1 Imperial bushel = 1.032 Winchester bushels

Same relations as above maintain for gallons (dry measure)

[NOTE.—1 U. S. gallon (dry) = 1.164 U. S. gallons (liquid).]

U. S. Units

2 pints	= 1 quart	=	67.2 cubic inches
4 quarts	= 1 gallon * = 8 pints	=	268.8 cubic inches
2 gallons *	= 1 peck = 16 pints = 8 quarts	=	537.6 cubic inches
4 pecks	= 1 bushel = 64 pints = 32 quarts = 8 gallons *	=	2150.42 cubic inches

1 cubic foot contains 6.428 gallons (dry measure) *

Liquid Measure

One U. S. gallon (liquid measure) contains 231 cubic inches. It holds 8.336 pounds distilled water at 62 deg fahr.

One British Imperial gallon contains 277.42 cubic inches. It holds 10 pounds distilled water at 62 deg fahr.

1 U. S. gallon (liquid) = 0.8327 Imperial gallon

1 Imperial gallon = 1.201 U. S. gallons (liquid)

[NOTE.—1 U. S. gallon (liquid) = 0.8594 U. S. gallon (dry).]

U. S. Units

4 gills	= 1 pint	=	16 fluid ounces
2 pints	= 1 quart = 8 gills	=	32 fluid ounces
4 quarts	= 1 gallon = 32 gills = 8 pints	=	128 fluid ounces

1 cubic foot contains 7.4805 gallons (liquid measure)

Apothecaries' Fluid Measure

60 minims = 1 fluid drachm. 8 drachms = 1 fluid ounce

In the U. S. a fluid ounce is the 128th part of a U. S. gallon, or 1.805 cu in. or 29.58 cu cm. It contains 455.8 grains of water at 62 deg fahr. In Great Britain the fluid ounce is 1.732 cu in. and contains 1 ounce avoirdupois (or 437.5 grains) of water at 62 deg. fahr.

* The *gallon* is not a U. S. legal dry measure.

Table 4. Plane Angle [*No Dimensions*]

Multiply Number of → by to Obtain ↓	Degrees	Minutes	Quadrants	Radians *	Revolutions * (Circumferences)	Seconds
Degrees	1	1.667×10^{-2}	90	57.30	360	2.778×10^{-4}
Minutes	60	1	5400	3438	2.16×10^{4}	1.667×10^{-2}
Quadrants	1.111×10^{-2}	1.852×10^{-4}	1	0.6366	4	3.087×10^{-6}
Radians *	1.745×10^{-2}	2.909×10^{-4}	1.571	1	6.283	4.848×10^{-6}
Revolutions * (Circumferences)	2.778×10^{-3}	4.630×10^{-5}	0.25	0.1591	1	7.716×10^{-7}
Seconds	3600	60	3.24×10^{5}	2.063×10^{5}	1.296×10^{6}	1

* 2π radians = 1 circumference = 360 degrees by definition.

Table 5. Solid Angle [*No Dimensions*]

Multiply Number of → by to Obtain ↓	Hemispheres	Spheres *	Spherical right angles	Steradians †
Hemispheres	1	2	0.25	0.1592
Spheres *	0.5	1	0.125	7.958×10^{-2}
Spherical right angles	4	8	1	0.6366
Steradians †	6.283	12.57	1.571	1

* A sphere is the total solid angle about a point. † 4π steradians = 1 sphere by definition.

Table 6. Time [T]

Multiply Number of → by to Obtain ↓	Days	Hours	Minutes	Months (average) *	Seconds	Weeks
Days	1	4.167×10^{-2}	6.944×10^{-4}	30.42	1.157×10^{-5}	7
Hours	24	1	1.667×10^{-2}	730.0	2.778×10^{-4}	168
Minutes	1440	60	1	4.380×10^{4}	1.667×10^{-2}	1.008×10^{4}
Months (average) *	3.288×10^{-2}	1.370×10^{-3}	2.283×10^{-5}	1	3.806×10^{-7}	0.2302
Seconds	8.64×10^{4}	3600	60	2.628×10^{6}	1	6.048×10^{5}
Weeks	0.1429	5.952×10^{-3}	9.921×10^{-5}	4.344	1.654×10^{-6}	1

* One common year = 365 days; one leap year = 366 days; one average month = 1/12 of a common year.

Table 7. Linear Velocity $[LT^{-1}]$

Multiply Number of → by / to Obtain ↓	Centimeters per second	Feet per minute	Feet per second	Kilometers per hour	Kilometers per minute	Knots*	Meters per minute	Meters per second	Miles per hour	Miles per minute
Centimeters per second	1	0.5080	30.48	27.78	1667	51.48	1.667	100	44.70	2682
Feet per minute	1.969	1	60	54.68	3281	101.3	3.281	196.8	88	5280
Feet per second	3.281×10^{-2}	1.667×10^{-2}	1	0.9113	54.68	1.689	5.468×10^{-2}	3.281	1.467	88
Kilometers per hour	0.036	1.829×10^{-2}	1.097	1	60	1.853	0.06	3.6	1.609	96.54
Kilometers per minute	0.0006	3.048×10^{-4}	1.829×10^{-2}	1.667×10^{-2}	1	3.088×10^{-2}	0.001	0.06	2.682×10^{-2}	1.609
Knots*	1.943×10^{-2}	9.868×10^{-3}	0.5921	0.5396	32.38	1	3.238×10^{-2}	1.943	0.8684	52.10
Meters per minute	0.6	0.3048	18.29	16.67	1000	30.88	1	60	26.82	1609
Meters per second	0.01	5.080×10^{-3}	0.3048	0.2778	16.67	0.5148	1.667×10^{-2}	1	0.4470	26.82
Miles per hour	2.237×10^{-2}	1.136×10^{-2}	0.6818	0.6214	37.28	1.152	3.728×10^{-2}	2.237	1	60
Miles per minute	3.728×10^{-4}	1.892×10^{-4}	1.136×10^{-2}	1.036×10^{-2}	0.6214	1.919×10^{-2}	6.214×10^{-4}	3.728×10^{-2}	1.667×10^{-2}	1

* Nautical miles per hour.

The Miner's Inch

(Used in Measuring Flow of Water)

An Act of the California legislature, May 23, 1901, makes the standard miner's inch 1.5 cu ft per minute, measured through any aperture or orifice.

The term miner's inch is more or less indefinite, for the reason that California water companies do not all use the same head above the center of the aperture, and the inch varies from 1.36 to 1.73 cu ft per minute, but the most common measurement is through an aperture 2 in. high and whatever length is required, and through a plank 1 1/4 in. thick. The lower edge of the aperture should be 2 in. above the bottom of the measuring-box, and the plank 5 in. high above the aperture, thus making a 6-in. head above the center of the stream. Each square inch of this opening represents a miner's inch, which is equal to a flow of 1.5 cu ft per minute.

Table 8. Angular Velocity $[T^{-1}]$

to Obtain ↓ \ Multiply Number of → by	Degrees per second	Radians per second	Revolutions per minute	Revolutions per second
Degrees per second	1	57.30	6	360
Radians per second	1.745×10^{-2}	1	0.1047	6.283
Revolutions per minute	0.1667	9.549	1	60
Revolutions per second	2.778×10^{-3}	0.1592	1.667×10^{-2}	1

Table 9. Linear Acceleration * $[LT^{-2}]$

to Obtain ↓ \ Multiply Number of → by	Centimeters per second per second	Feet per second per second	Kilometers per hour per second	Meters per second per second	Miles per hour per second
Centimeters per second per second	1	30.48	27.78	100	44.70
Feet per second per second	3.281×10^{-2}	1	0.9113	3.281	1.467
Kilometers per hour per second	0.036	1.097	1	3.6	1.609
Meters per second per second	0.01	0.3048	0.2778	1	0.4770
Miles per hour per second	2.237×10^{-2}	0.6818	0.6214	2.237	1

* The (standard) acceleration due to gravity (g_0) = 980.7 cm per sec per sec = 32.17 feet per sec per sec = 35.30 km per hour per sec = 9.807 meters per sec per sec = 21.94 miles per hour per sec.

Table 10. Angular Acceleration $[T^{-2}]$

to Obtain ↓ \ Multiply Number of → by	Radians per second per second	Revolutions per minute per minute	Revolutions per minute per second	Revolutions per second per second
Radians per second per second	1	1.745×10^{-3}	0.1047	6.283
Revolutions per minute per minute	573.0	1	60	3600
Revolutions per minute per second	9.549	1.667×10^{-2}	1	60
Revolutions per second per second	0.1592	2.778×10^{-4}	1.667×10^{-2}	1

Table 11. Mass [M] and Weight *

Multiply Number of → by to Obtain ↓	Grains	Grams	Kilograms	Milligrams	Ounces †	Pounds †	Tons (long)	Tons (metric)	Tons (short)
Grains	1	15.43	1.543×10^{4}	1.543×10^{-2}	437.5	7000			
Grams	6.481×10^{-2}	1	1000	0.001	28.35	453.6	1.016×10^{6}	10^{6}	9.072 10^{5}
Kilograms	6.481×10^{-5}	0.001	1	10^{-6}	$2\ 835 \times 10^{-2}$	0 4536	1016	1000	907.2
Milligrams	64.81	1000	10^{6}	1	2.835×10^{4}	4.536×10^{5}	1.016×10^{9}	10^{9}	$9\ 072 \times 10^{8}$
Ounces †	2.286×10^{-3}	3.527×10^{-2}	35.27	3.527×10^{-5}	1	16	3.584×10^{4}	3.527×10^{4}	3.2×10^{4}
Pounds †	1.429×10^{-4}	2.205×10^{-3}	2.205	2.205×10^{-6}	6.250×10^{-2}	1	2240	2205	2000
Tons (long)		9.842×10^{-7}	9.842×10^{-4}	9.842×10^{-10}	2.790×10^{-5}	4.464×10^{-4}	1	0.9842	0.8929
Tons (metric)		10^{-6}	0.001	10^{-9}	2.835×10^{-5}	$4\ 536 \times 10^{-4}$	1.016	1	0.9072
Tons (short)		1.102×10^{-6}	1.102×10^{-3}	$1\ 102 \times 10^{-9}$	3.125×10^{-5}	0.0005	1.120	1.102	1

* These same conversion factors apply to the *gravitational* units of force having the corresponding names. The dimensions of these units when used as gravitational units of force are MLT^{-2}; see table for *Force*.

† Avoirdupois pounds and ounces.

Metric Multiples

10^6 micrograms = 10^3 milligrams = 10^2 centigrams = 10 decigrams = 1 gram = 10^{-1} dekagram = 10^{-2} hectogram = 10^{-3} kilogram = 10^{-4} myriagram = 10^{-6} megagram

Avoirdupois Weight

(Used Commercially)

27.343 grains = 1 drachm
16 drachms = 1 ounce (oz) = 437.5 grains
16 ounces = 1 pound (lb) = 7000 grains
28 pounds = 1 quarter (qr)
4 quarters = 1 hundredweight (cwt) = 112 pounds
20 hundredweight = 1 gross or long ton *
2000 pounds = 1 net or short ton

(* NOTE.—The long ton is used by the U. S. custom-houses in collecting duties upon foreign goods. It is also used in freighting coal and selling it wholesale.)

14 pounds = 1 stone; 100 pounds = 1 quintal

Troy Weight

(Used in weighing gold or silver)

24 grains = 1 pennyweight (dwt)
20 pennyweights = 1 ounce (oz) = 480 grains
12 ounces = 1 pound (lb) = 5760 grains

The grain is the same in Avoirdupois, Troy and Apothecaries' weights. A carat, for weighing diamonds = 3.086 grains = 0.200 gram. (International Standard, 1913.)

1 pound troy = .8229 pound avoirdupois
1 pound avoirdupois = 1.2153 pounds troy

Apothecaries' Weight

(Used in compounding medicines)

20 grains = 1 scruple (℈)
3 scruples = 1 drachm (ʒ) = 60 grains
8 drachms = 1 ounce (℥) = 480 grains
12 ounces = 1 pound (lb) = 5760 grains

The grain is the same in Avoirdupois, Troy and Apothecaries' weights.

1 pound apothecaries = 0.82286 pound avoirdupois
1 pound avoirdupois = 1.2153 pounds apothecaries

Table 12. Density or Mass per Unit Volume $[ML^{-3}]$

Multiply Number of → by to Obtain ↓	Grams per cubic centimeter	Kilograms per cubic meter	Pounds per cubic foot	Pounds per cubic inch
Grams per cubic centimeter	1	0.001	1.602×10^{-2}	27.68
Kilograms per cubic meter	1000	1	16.02	2.768×10^{4}
Pounds per cubic foot	62.43	6.243×10^{-2}	1	1728
Pounds per cubic inch	3.613×10^{-2}	3.613×10^{-5}	5.787×10^{-4}	1
Pounds per mil foot *	3.405×10^{-7}	3.405×10^{-10}	5.456×10^{-9}	9.425×10^{-6}

* Unit of volume is a volume one foot long and one circular mil in cross-section area.

Table 13. Force * $[MLT^{-2}]$ **or** $[F]$

Multiply Number of → by to Obtain ↓	Dynes	Grams	Joules per cm	Joules per meter (newtons)	Kilo-grams	Pounds	Poundals
Dynes	1	980.7	10^{7}	10^{5}	9.807×10^{5}	4.448×10^{5}	1.383×10^{4}
Grams	1.020×10^{-3}	1	1.020×10^{4}	102.0	1000	453.6	14.10
Joules per cm	10^{-7}	9.807×10^{-5}	1	.01	9.807×10^{-2}	4.448×10^{-2}	1.383×10^{-3}
Joules per meter (newtons)	10^{-5}	9.807×10^{-3}	100	1	9.807	4.448	0.1383
Kilograms	1.020×10^{-6}	0.001	10.20	0.1020	1	0.4536	1.410×10^{-2}
Pounds	2.248×10^{-6}	2.205×10^{-3}	22.48	0.2248	2.205	1	3.108×10^{-2}
Poundals	7.233×10^{-5}	7.093×10^{-2}	723.3	7.233	70.93	32.17	1

* Conversion factors between absolute and gravitational units apply only under standard acceleration due to gravity conditions.

Table 14. Torque or Moment of Force $[ML^2T^{-2}]$ or $[FL]$ *

Multiply Number of → by to Obtain ↓	Dyne-centimeters	Gram-centimeters	Kilogram-meters	Pound-feet	Newton-meter
Dyne-centimeters	1	980.7	9.807×10^7	1.356×10^7	10^7
Gram-centimeters	1.020×10^{-3}	1	10^5	1.383×10^4	1.020×10^4
Kilogram-meters	1.020×10^{-8}	10^{-5}	1	0.1383	0.1020
Pound-feet	7.376×10^{-8}	7.233×10^{-5}	7.233	1	0.7376
Newton-meter	10^{-7}	9.807×10^{-4}	9.807	1.305	1

* Same dimensions as energy.

Table 15. Pressure or Force per Unit Area $[ML^{-1}T^{-2}]$ or $[FL^{-2}]$

Multiply Number of → by to Obtain ↓	Atmospheres *	Baryes or dynes per square centimeter †	Centimeters of mercury at 0° C ‡	Inches of mercury at 0° C ‡	Inches of water at 4° C	Kilograms per square meter §	Pounds per square foot	Pounds per square inch	Tons (short) per square foot	Newtons per square meter
Atmospheres *	1	9.869×10^{-7}	1.316×10^{-2}	3.342×10^{-2}	2.458×10^{-3}	9.678×10^{-5}	4.725×10^{-4}	6.804×10^{-2}	0.9450	9.869×10^{-6}
Baryes or dynes per square centimeter †	1.013×10^6	1	1.333×10^4	3.386×10^4	2.491×10^{-3}	98.07	478.8	6.895×10^4	9.576×10^5	10
Centimeters of mercury at 0° C ‡	76.00	7.501×10^{-5}	1	2.540	0.1868	7.356×10^{-3}	3.591×10^{-2}	5.171	71.83	7.501×10^{-4}
Inches of mercury at 0° C ‡	29.92	2.953×10^{-5}	0.3937	1	7.355×10^{-2}	2.896×10^{-3}	1.414×10^{-2}	2.036	28.28	2.953×10^{-4}
Inches of water at 4° C	406.8	4.015×10^{-4}	5.354	13.60	1	3.937×10^{-2}	0.1922	27.68	384.5	4.015×10^{-3}
Kilograms per square meter §	1.033×10^4	1.020×10^{-2}	136.0	345.3	25.40	1	4.882	703.1	9765	0.1020
Pounds per square foot	2117	2.089×10^{-3}	27.85	70.73	5.204	0.2048	1	144	2000	2.089×10^{-2}
Pounds per square inch	14.70	1.450×10^{-5}	0.1934	0.4912	3.613×10^{-2}	1.422×10^{-3}	6.944×10^{-3}	1	13.89	1.450×10^{-4}
Tons (short) per square foot	1.058	1.044×10^{-6}	1.392×10^{-2}	3.536×10^{-2}	2.601×10^{-3}	1.024×10^{-4}	0.0005	0.072	1	1.044×10^{-5}
Newtons per square meter	1.013×10^5	10^{-1}	1.333×10^3	3.386×10^3	2.491×10^{-4}	9.807	47.88	6.895×10^3	9.576×10^4	1

* Definition: One atmosphere (standard) = 76 cm of mercury at 0 deg cent.
† Sometimes called a bar.
‡ To convert height h of a column of mercury at t degrees Centigrade to the equivalent height h_0 at 0 deg cent use $h_0 = h\left\{1 - \frac{(m-l)t}{1+mt}\right\}$ where $m = 0.0001818$ and $l = 18.4 \times 10^{-6}$ if the scale is engraved on brass; $l = 8.5 \times 10^{-6}$ if on glass. This assumes the scale is correct at 0 deg cent; for other cases (any liquid) see *International Critical Tables*, Vol. 1, 68.
§ 1 gram per sq cm = 10 kilograms per sq m.

Table 16. Energy, Work and Heat * [ML^2T^{-2}] or [FL]

Multiply Number of → by to Obtain ↓	British thermal units †	Centimeter-grams	Ergs or centimeter-dynes	Foot-pounds	Horsepower-hours	Joules ‡ or watt-seconds	Kilogram-calories †	Kilowatt-hours	Meter-kilograms	Watt-hours
British thermal units †	1	9.297×10^{-8}	9.480×10^{-11}	1.285×10^{-3}	2545	9.480×10^{-4}	3.969	3413	9.297×10^{-3}	3.413
Centimeter-grams	1.076×10^{7}	1	1.020×10^{-3}	1.383×10^{4}	2.737×10^{10}	1.020×10^{4}	4.269×10^{7}	3.671×10^{10}	10^{5}	3.671×10^{7}
Ergs or centimeter-dynes	1.055×10^{10}	980.7	1	1.356×10^{7}	2.684×10^{13}	10^{7}	4.186×10^{10}	3.6×10^{13}	9.807×10^{7}	3.6×10^{10}
Foot-pounds	778.0	7.233×10^{-5}	7.367×10^{-8}	1	1.98×10^{6}	0.7376	3087	2.655×10^{6}	7.233	2655
Horsepower-hours	3.929×10^{-4}	3.654×10^{-11}	3.722×10^{-14}	5.050×10^{-7}	1	3.722×10^{-7}	1.559×10^{-3}	1.341	3.653×10^{-6}	1.341×10^{-3}
Joules ‡ or watt-seconds	1054.8	9.807×10^{-5}	10^{-7}	1.356	2.684×10^{6}	1	4186	3.6×10^{6}	9.807	3600
Kilogram-calories †	0.2520	2.343×10^{-8}	2.389×10^{-11}	3.239×10^{-4}	641.3	2.389×10^{-4}	1	860.0	2.343×10^{-3}	0.8600
Kilowatt-hours	2.930×10^{-4}	2.724×10^{-11}	2.778×10^{-14}	3.766×10^{-7}	0.7457	2.778×10^{-7}	1.163×10^{-3}	1	2.724×10^{-6}	0.001
Meter-kilograms	107.6	10^{-5}	1.020×10^{-8}	0.1383	2.737×10^{5}	0.1020	426.9	3.671×10^{5}	1	367.1
Watt-hours	0.2930	2.724×10^{-8}	2.778×10^{-11}	3.766×10^{-4}	745.7	2.778×10^{-4}	1.163	1000	2.724×10^{-3}	1

* See note at the bottom of Table 17.

† Mean calorie and Btu used throughout. One gram-calorie = 0.001 kilogram-calorie; one Ostwald calorie = 0.01 kilogram-calorie.

The IT cal, 1000 international steam-table calories, has been defined as the 1/860th part of the international kilowatthour (see *Mechanical Engineering*, Nov., 1935, p. 710). Its value is very nearly equal to the mean kilogram-calorie, 1 IT cal = 1.00037 kilogram-calories (mean). 1 Btu = 251.996 IT cal.

‡ Absolute joule, defined as 10^7 ergs. The international joule, based on the international ohm and ampere, equals 1.0003 absolute joules.

Table 17. Power or Rate of Doing Work * $[ML^2T^{-3}]$ or $[FLT^{-1}]$

Multiply Number of → by to Obtain ↓	British thermal units per minute	Ergs per second	Foot-pounds per minute	Foot-pounds per second	Horsepower *	Kilogram-calories per minute	Kilowatts	Metric horsepower	Watts
British thermal units per minute	1	5.689×10^{-9}	1.285×10^{-3}	7.712×10^{-2}	42.41	3.969	56.89	41.83	5.689×10^{-2}
Ergs per second	1.758×10^{8}	1	2.259×10^{5}	1.356×10^{7}	7.457×10^{9}	6.977×10^{8}	10^{10}	7.355×10^{9}	10^{7}
Foot-pounds per minute	778.0	4.426×10^{-6}	1	60	3.3×10^{4}	3087	4.426×10^{4}	3.255×10^{4}	44.26
Foot-pounds per second	12.97	7.376×10^{-8}	1.667×10^{-2}	1	550	51.44	737.6	542.5	0.7376
Horsepower *	2.357×10^{-2}	1.341×10^{-10}	3.030×10^{-5}	1.818×10^{-3}	1	9.355×10^{-2}	1.341	0.9863	1.341×10^{-3}
Kilogram-calories per minute	0.2520	1.433×10^{-9}	3.239×10^{-4}	1.943×10^{-2}	10.69	1	14.33	10.54	1.433×10^{-2}
Kilowatts	1.758×10^{-2}	10^{-10}	2.260×10^{-5}	1.356×10^{-3}	0.7457	$6\ 977 \times 10^{-2}$	1	0.7355	10^{-3}
Metric horsepower	2.390×10^{-2}	1.360×10^{-10}	3.072×10^{-5}	1.843×10^{-3}	1.014	9.485×10^{-2}	1.360	1	1.360×10^{-3}
Watts	17.58	10^{-7}	2.260×10^{-2}	1.356	745.7	69.77	1000	735.5	1

1 Cheval-vapeur = 75 kilogram-meters per second
1 Poncelet = 100 kilogram-meters per second

* The "horsepower" used in these tables is equal to 550 foot-pounds per second by definition. Other definitions are one horsepower equals 746 watts (U. S. and Great Britain) and one horsepower equals 736 watts (continental Europe). Neither of these latter definitions is equivalent to the first; the "horsepowers" defined in these latter definitions are widely used in the rating of electrical machinery.

Table 18. Quantity of Electricity and Electric Flux $[Q]$

Multiply Number of → by to Obtain ↓	Abcoulombs	Ampere-hours	Coulombs	Faradays	Stat-coulombs
Abcoulombs *	1	360	0.1	9649	3.335×10^{-11}
Ampere-hours	2.778×10^{-3}	1	2.778×10^{-4}	26.80	9.259×10^{-14}
Coulombs	10	3600	1	9.649×10^{4}	3.335×10^{-10}
Faradays	1.036×10^{-4}	3.731×10^{-2}	1.036×10^{-5}	1	3.457×10^{-15}
Statcoulombs *	2.998×10^{10}	1.080×10^{13}	2.998×10^{9}	2.893×10^{14}	1

* Conventionally, in the electrostatic and electromagnetic systems of units, the number of lines of electric flux emanating from a point charge is 4π times that charge (or quantity of electricity). The statcoulomb and the abcoulomb are units of charge, not flux.

Table 19. Charge per Unit Area and Electric Flux Density $[QL^{-2}]$

Multiply Number of → by to Obtain ↓	Abcoulombs per square centimeter *	Coulombs per square centimeter	Coulombs per square inch	Statcoulombs per square centimeter	Coulombs per square meter
Abcoulombs per square centi-centimeter *	1	0.1	1.550×10^{-2}	3.335×10^{-11}	10^{-5}
Coulombs per square centimeter	10	1	0.1550	3.335×10^{-10}	10^{-4}
Coulombs per square inch	64.52	6.452	1	2.151×10^{-9}	6.452×10^{-4}
Statcoulombs per square centimeter *	2.998×10^{10}	2.998×10^{9}	4.647×10^{8}	1	2.998×10^{5}
Coulombs per square meter	10^{5}	10^{4}	1550	3.335×10^{-6}	1

* See footnote to Table 18.

Table 20. Electric Current $[QT^{-1}]$

Multiply Number of → by to Obtain ↓	Abamperes	Amperes	Statamperes
Abamperes	1	0.1	3.335×10^{-11}
Amperes	10	1	3.335×10^{-10}
Statamperes	2.998×10^{10}	2.998×10^{9}	1

Table 21. Current Density $[QT^{-1}L^{-2}]$

Multiply Number of → by to Obtain ↓	Abamperes per square centimeter	Amperes per square centimeter	Amperes per square inch	Statamperes per square centimeter	Amperes per square meter
Abamperes per square centimeter	1	0.1	1.550×10^{-2}	3.335×10^{-11}	10^{-5}
Amperes per square centimeter	10	1	0.1550	3.335×10^{-10}	10^{-4}
Amperes per square inch	64.52	6.452	1	2.151×10^{-9}	6.452×10^{-4}
Statamperes per square centimeter	2.998×10^{10}	2.998×10^{9}	4.647×10^{8}	1	2.998×10^{5}
Amperes per square meter	10^{5}	10^{4}	1550	3.335×10^{-6}	1

Table 22. Electric Potential and Electromotive Force $[MQ^{-1}L^2T^{-2}]$ or $[FQ^{-1}L]$

To Obtain ↓ by Multiply Number of →	Abvolts	Microvolts	Millivolts	Statvolts	Volts
Abvolts	1	100	10^5	2.998×10^{10}	10^8
Microvolts	0.01	1	1000	2.998×10^8	10^6
Millivolts	10^{-5}	0.001	1	2.998×10^5	1000
Statvolts	3.335×10^{-11}	3.335×10^{-9}	3.335×10^{-6}	1	3.335×10^{-3}
Volts	10^{-8}	10^{-6}	0.001	299.8	1

Table 23. Electric Field Intensity and Potential Gradient $[MQ^{-1}LT^{-2}]$ or $[FQ^{-1}]$

To Obtain ↓ by Multiply Number of →	Abvolts per centimeter	Microvolts per meter	Millivolts per meter	Statvolts per centimeter	Volts per centimeter	Kilovolts per centimeter	Volts per inch	Volts per mil	Volts per meter
Abvolts per centimeter	1	1	1000	2.998×10^{10}	10^8	10^{11}	3.937×10^7	3.937×10^{10}	10^6
Microvolts per meter	1	1	1000	2.998×10^{10}	10^8	10^{11}	3.937×10^7	3.937×10^{10}	10^6
Millivolts per meter	0.001	0.001	1	2.998×10^7	10^5	10^8	3.937×10^4	3.937×10^7	1000
Statvolts per centimeter	3.335×10^{-11}	3.335×10^{-11}	3.335×10^{-8}	1	3.335×10^{-3}	3.335	1.313×10^{-3}	1.313	3.335×10^{-5}
Volts per centimeter	10^{-8}	10^{-8}	10^{-5}	299.8	1	1000	0.3937	393.7	10^{-2}
Kilovolts per centimeter	10^{-11}	10^{-11}	10^{-8}	0.2998	0.001	1	3.937×10^{-4}	0.3937	10^{-5}
Volts per inch	2.540×10^{-8}	2.540×10^{-8}	2.540×10^{-5}	761.6	2.540	2540	1	1000	2.540×10^{-2}
Volts per mil	2.540×10^{-11}	2.540×10^{-11}	2.540×10^{-8}	0.7616	2.540×10^{-3}	2.540	0.001	1	2.540×10^{-5}
Volts per meter	10^{-6}	10^{-6}	10^{-3}	2.998×10^4	100	10^5	39.37	3.937×10^4	1

Table 24. Electric Resistance $[MQ^{-2}L^{2}T^{-1}]$ or $[FQ^{-2}LT]$

Multiply Number of → by, to Obtain ↓	Abohms	Megohms	Microhms	Ohms	Statohms
Abohms	1	10^{15}	1000	10^{9}	8.988×10^{20}
Megohms	10^{-15}	1	10^{-12}	10^{-6}	8.988×10^{5}
Microhms	0.001	10^{12}	1	10^{6}	8.988×10^{17}
Ohms	10^{-9}	10^{6}	10^{-6}	1	8.988×10^{11}
Statohms	1.112×10^{-21}	1.112×10^{-6}	1.112×10^{-18}	1.112×10^{-12}	1

Electrical Conductance $[F^{-1}QL^{-1}T^{-1}]$

1 mho = 1 ohm^{-1} = 10^{-6} megmho = 10^{6} micromho

Table 25. Electric Resistivity * $[MQ^{-2}L^{3}T^{-1}]$ or $[FQ^{-2}L^{2}T]$

Multiply Number of → by, to Obtain ↓	Abohm-centimeters	Microhm-centi-meters	Microhm-inches	Ohms (mil, foot)	Ohms (meter, gram) †	Ohm-meters
Abohm-centimeters	1	1000	2540	166.2	$\frac{10^{5}}{\delta}$	10^{11}
Microhm-centimeters	0.001	1	2.540	0.1662	$\frac{100}{\delta}$	10^{8}
Microhm-inches	3.937×10^{-4}	0.3937	1	6.545×10^{-2}	$\frac{39.37}{\delta}$	3.937×10^{7}
Ohms (mil, foot)	6.015×10^{-3}	6.015	15.28	1	$\frac{601.5}{\delta}$	6.015×10^{8}
Ohms (meter, gram) †	$10^{-5}\delta$	0.01δ	$2.540 \times 10^{-2}\delta$	$1.662 \times 10^{-3}\delta$	1	$10^{-6}\delta$
Ohm-meters	10^{-11}	10^{-8}	2.540×10^{-8}	1.662×10^{-9}	$\frac{10^{-6}}{\delta}$	1

* In this table δ is density in grams per cm.^{3} The following names, corresponding respectively to those at the tops of columns, are sometimes used: abohms per cm cube; microhms per cm cube; microhms per inch cube; ohms per mil-foot; ohms per meter-gram. The first four columns are headed by units of *volume* resistivity, the last by a unit of *mass* resistivity. The dimensions of the latter are $Q^{-2}L^{6}T^{-1}$; not these given in the heading of the table.

† One ohm (meter, gram) = 5710 ohms (mile, pound).

Table 26. Electric Conductivity * $[M^{-1}Q^2L^{-3}T]$ or $[F^{-1}Q^2L^{-2}T^{-1}]$

Multiply Number of → by to Obtain ↓	Abmhos per cm	Mhos (mil, foot)	Mhos (meter, gram)	Micro-mhos per cm	Micro-mhos per inch	Mhos per meter
Abmhos per cm	1	6.015×10^{-3}	$10^{-5}\delta$	0.001	3.937×10^{-4}	10^{-11}
Mhos (mil, foot)	166.2	1	$1.662 \times 10^{-3}\delta$	0.1662	6.524×10^{-2}	1.662×10^{-9}
Mhos (meter, gram)	$10^5/\delta$	$601.5/\delta$	1	$100/\delta$	$39.37/\delta$	$10^{-6}/\delta$
Micromhos per cm	1000	6.015	0.01δ	1	0.3937	10^{-8}
Micromhos per inch	2540	15.28	$2.540 \times 10^{-2}\delta$	2.540	1	2.54×10^{-8}
Mhos per meter	10^{11}	6.015×10^8	$10^6\delta$	10^8	3.937×10^7	1

* See footnote of Table 25, Electric Resistivity. Names sometimes used are abmho per cm cube, mho per mil-foot, etc. Dimensions of mass conductivity are $Q^2L^{-6}T$.

Table 27. Capacitance $[M^{-1}Q^2L^{-2}T^2]$ or $[F^{-1}Q^2L^{-1}]$

Multiply Number of → by to Obtain ↓	Abfarads	Farads	Microfarads	Statfarads
Abfarads	1	10^{-9}	10^{-15}	1.112×10^{-21}
Farads	10^9	1	10^{-6}	1.112×10^{-12}
Microfarads	10^{15}	10^6	1	1.112×10^{-6}
Statfarads	8.988×10^{20}	8.988×10^{11}	8.988×10^5	1

Table 28. Inductance $[MQ^{-2}L^2]$ or $[FQ^{-2}LT^2]$

to Obtain ↓ / by / Multiply Number of →	Abhenrys *	Henrys	Microhenrys	Millihenrys	Stathenrys
Abhenrys *	1	10^9	1000	10^6	8.988×10^{20}
Henrys	10^{-9}	1	10^{-6}	0.001	8.988×10^{11}
Microhenrys	0.001	10^6	1	1000	8.988×10^{17}
Millihenrys	10^{-6}	1000	0.001	1	8.988×10^{14}
Stathenrys	1.112×10^{-21}	1.112×10^{-12}	1.112×10^{-18}	1.112×10^{-15}	1

* An abhenry is sometimes called a "centimeter." See footnote to Table 30 on "Magnetic Flux Density."

Table 29. Magnetic Flux $[MQ^{-1}L^2T^{-1}]$ or $[FQ^{-1}LT]$

to Obtain ↓ / by / Multiply Number of →	Kilolines	Maxwells (or lines)	Webers
Kilolines	1	0.001	10^5
Maxwells (or lines)	1000	1	10^8
Webers	10^{-5}	10^{-8}	1

Table 30. Magnetic Flux Density $[MQ^{-1}T^{-1}]$ or $[FQ^{-1}L^{-1}T]$

to Obtain ↓ / by / Multiply Number of →	Gausses (or lines per square centimeter)	Lines per square inch	Webers per square centimeter	Webers per square inch	Webers per square meter
Gausses (or lines per square centimeter)	1	0.1550	10^8	1.550×10^7	10^4
Lines per square inch	6.452	1	6.452×10^8	10^8	6.452×10^4
Webers per square centimeter	10^{-8}	1.550×10^{-9}	1	0.1550	10^{-4}
Webers per square inch	6.452×10^{-8}	10^{-8}	6.452	1	6.452×10^{-4}
Webers per square meter	10^{-4}	1.550×10^{-5}	10^4	1550	1

Table 31. Magnetic Potential and Magnetomotive Force $[QT^{-1}]$

to Obtain ↓ by Multiply Number of →	Abampere-turns	Ampere-turns	Gilberts
Abampere-turns	1	0.1	7.958×10^{-2}
Ampere-turns	10	1	0.7958
Gilberts	12.57	1.257	1

Table 32. Magnetic Field Intensity, Potential Gradient, and Magnetizing Force $[QL^{-1}T^{-1}]$

to Obtain ↓ by Multiply Number of →	Abampere-turns per centimeter	Ampere-turns per centimeter	Ampere-turns per inch	Oersteds (gilberts per centimeter)	Ampere-turns per meter
Abampere-turns per centimeter	1	0.1	3.937×10^{-2}	7.958×10^{-2}	10^{-3}
Ampere-turns per centimeter	10	1	0.3937	0.7958	10^{-2}
Ampere-turns per inch	25.40	2.540	1	2.021	2.54×10^{-2}
Oersteds (gilberts per centimeter)	12.57	1.257	0.4950	1	1.257×10^{-2}
Ampere-turns per meter	10^3	10^2	39.37	79.58	1

Table 33. Specific Heat $[L^2T^{-2}t^{-1}]$

(t = temperature)

To convert specific heat in any unit given to any other unit multiply the number of original units by a factor obtained by dividing the factor in the last column for the final unit by the factor for the original unit.

Unit of Heat or Energy	Unit of Mass	Temperature Scale*	Factor
Gram-calories	Gram	Centigrade	1
Kilogram-calories	Kilogram	Centigrade	1
British thermal units	Pound	Centigrade	1.800
British thermal units	Pound	Fahrenheit	1.000
Joules	Gram	Centigrade	4.186
Joules	Pound	Fahrenheit	1055.
Kilowatt-hours	Kilogram	Centigrade	1.163×10^{-3}
Kilowatt-hours	Pound	Fahrenheit	2.930×10^{-4}

* Temperature conversion formulas:

t_c = temperature in Centigrade degrees
t_f = temperature in Fahrenheit degrees
1 deg fahr = (5/9) deg cent.
$t_c = \frac{5}{9}(t_f - 32)$
$t_f = \frac{9}{5}t_c + 32$

Table 34. Thermal Conductivity $[MLT^{-3}t^{-1}]$ and Thermal Resistivity $[M^{-1}L^{-1}T^{3}t]$

(t = temperature)

To convert thermal conductivity, in gram-calories transmitted per second from one face of a cube 1 cm on edge to the opposite face per degree centigrade temperature difference between these faces, to the units given in any line of the following table, multiply by the factor in the last column.

To convert thermal conductivity in any unit given to any other unit multiply the number of original units by a factor obtained by dividing the factor in the last column for the final unit by the factor for the original unit.

To convert thermal resistivity, in degrees centigrade between one face of a cube 1 cm on edge and the opposite face per gram-calories transmitted per second between these faces, to the units given in any line of the following table, divide by the factor in the last column.

To convert thermal resistivity in any given unit to any other unit multiply the number of the original units by a factor obtained by dividing the factor in the last column for the original unit by the factor for the final unit.

Surface emission resistance in thermal ohms per square centimeter is derived from degrees fahrenheit per Btu per hour per square foot by multiplying the number of the latter units by 1761.

Units of				Temperature Scale	Factor
Heat	Area	Thickness	Time		
Gram-calories	cm^2	cm	second	Centigrade	1
Kilogram-calories	m^2	cm	hour	Centigrade	3.6×10^4
British thermal units	ft^2	inch	hour	Fahrenheit	2903.
Joules *	cm^2	cm	second	Centigrade	4.186
Joules	ft^2	inch	second	Fahrenheit	850.6
Kilowatt-hours	m^2	cm	hour	Centigrade	41.86
Kilowatt-hours	ft^2	inch	hour	Fahrenheit	0.8506

* Thermal resistances in these units are known as *thermal ohms.*

Table 35. Light

Multiply Number of → by to Obtain ↓	International candles	Hefners	10-cp pentanes	Carcels	Bougie decimales	English candles	German candles
International candles	1.00	0.90	10.0	9.61	1.00	1.04	1.055
Hefners	1.11	1.00	11.1	10.66	1.11	1.154	1.17
10-cp pentanes	0.10	0.09	1.00	0.96	0.10	0.104	0.105
Carcels	0.104	0.094	1.04	1.00	0.104	0.1	0.109
Bougie decimales	1.00	0.90	10.0	9.61	1.00	1.04	1.055
English candles	0.96	0.864	9.6	9.24	0.96	1.00	1.02
German candles	0.95	0.855	9.5	9.19	0.95	0.98	1.00

18. GAGES

SHEET METAL GAGES. The important sheet metal gages in use in the United States are: the United States Standard Gage for sheet and plate iron and steel, the American Wire Gage (also called the Brown and Sharpe W.G.) for copper, aluminum, and brass and other non-ferrous alloys, the Tin Plate Gage, the Galvanized Sheet Gage, the American Zinc Gage, and the Birmingham Wire (or Stubs' Iron Wire) Gage. In Canada and England the Birmingham Gage (different from the Birmingham Wire Gage) and the Imperial Standard Wire Gage (S.W.G.) are used. Still other gages are used elsewhere. In Japan standard thickness of sheet metal is denoted by the thickness in millimeters. A standard Decimal Gage, in which the standard thicknesses are denoted by decimal parts of an inch and not by gage numbers, has been used in the United States. Copper sheets may be obtained with thicknesses any integral multiple of $^1/_{16}$ of an inch up to 2 in. Heavy copper sheets may be obtained in definite weights per square foot. Each ounce of weight is equivalent to approximately 0.001352 in. thickness. Lead is usually ordered in this manner, each pound being equivalent to approximately 0.017 in. thickness.

The United States Standard Gage for sheet iron and steel (Act of Congress, March 3, 1893; formerly the legal standard for duties) is a *weight* gage based on a density for wrought iron of 480 pounds per cubic foot. Since 1893, steel (density of 489.6 lb per cu ft) has come into general use. A given gage number of this gage represents a fixed weight per unit area; hence a steel sheet will have a smaller thickness than a wrought iron sheet of the same gage number. Monel metal sheets are rolled to the thickness given for wrought iron without regard to its weight, which is about 552.2 lb per cu ft. Practice among steel manufacturers is irregular, some keeping the *thickness* constant for a given gage number irrespective of weight. If this practice is followed, the weight per square foot and per square meter given in the second and third columns of Table 36 will vary, whereas thickness will remain near that given for wrought iron.

The American Wire Gage specifies thicknesses without regard to weight. For the basis of this gage see Wire Gages, p. 1-70, where are also given the Birmingham W.G and the S.W.G.

Tables of Thickness and Weight corresponding to United States Standard gage and American Wire gage numbers are shown in Tables 36 and 37. These tables are taken from *Circular* 391 of the Bureau of Standards, in which are given all the gages mentioned above and the tolerances customary in commerce.

WIRE GAGES. The sizes of wires having a diameter less than $^1/_2$ in. are usually stated in terms of certain arbitrary scales called "gages." The size or gage number of a solid wire refers to the cross-section of the wire perpendicular to its length; the size or gage number of a stranded wire refers to the total cross-section of the constituent wires, irrespective of the pitch of the spiraling. Larger wires are usually described in terms of their area expressed in circular mils. A circular mil is the area of a circle 1 mil in diameter, and the area of any circle in circular mils is equal to the square of its diameter in mils.

There are a number of wire gages in use, the principal ones being the following:

American or Brown and Sharpe Wire Gage. This gage is the one commonly used in the United States for copper, aluminum, and resistance wires. The gage is designated by either of the abbreviations A.W.G. or B. & S.

Basis of the A.W.G. or B. & S. Gage. The diameters of wires having successive numbers on this gage are in the ratio of $\sqrt[39]{92}$(= 1.1229 approx.) to 1, and the No. 36 wire has a diameter of 5 mils. No. 35 A.W.G., therefore, has a diameter of $5 \times 1.1229 = 5.61$ mils, and so on until No. 0000 is reached, having a diameter of 460 mils.

The ratio $\sqrt[39]{92}$ is approximately equal to $\sqrt[6]{2}$, which is 1.1225. This circumstance makes it possible to have a group of wires of regular gage size with an aggregate area approximately equal to that of another regular gage size. For example, a reduction of three gage numbers (as from gage No. 36 to No. 33) results in a new gage number representing a diameter approximately $\sqrt{2}$ times that represented by the original gage number—or an area approximately two times as great.

The following approximate relations are also useful:

An increase of 1 in the number increases the resistance 25 per cent.
An increase of 2 in the number increases the resistance 60 per cent.
An increase of 3 in the number increases the resistance 100 per cent.
An increase of 10 in the number increases the resistance 10 times.

Table 36. United States Standard Gage * for Sheet and Plate Iron and Steel, and Its Extension †

Gage No.	Weight per square foot		Weight per square meter	Approximate thickness			
				Wrought iron 480 lb/ft^3		Steel and open-hearth iron 489.6 lb/ft^3	
	Ounces	Pounds	kg	Inch	mm	Inch	mm
0000000	320	20.00	97.65	0.500	12.70	0.490	12.45
000000	300	18.75	91.55	.469	11.91	.460	11.67
00000	280	17.50	85.44	.438	11.11	.429	10.90
0000	260	16.25	79.34	.406	10.32	.398	10.12
000	240	15.00	73.24	.375	9.52	.368	9.34
00	220	13.75	67.13	.344	8.73	.337	8.56
0	200	12.50	61.03	.312	7.94	.306	7.78
1	180	11.25	54.93	.2812	7.14	.2757	7.00
2	170	10.62	51.88	.2656	6.75	.2604	6.62
3	160	10.00	48.82	.2500	6.35	.2451	6.23
4	150	9.375	45.77	.2344	5.95	.2298	5.84
5	140	8.750	42.72	.2188	5.56	.2145	5.45
6	130	8.125	39.67	.2031	5.16	.1991	5.06
7	120	7.500	36.62	.1875	4.76	.1838	4.67
8	110	6.875	33.57	.1719	4.37	.1685	4.28
9	100	6.250	30.52	.1562	3.97	.1532	3.89
10	90	5.625	27.46	.1406	3.57	.1379	3.50
11	80	5.000	24.41	.1250	3.18	.1225	3.11
12	70	4.375	21.36	.1094	2.778	.1072	2.724
13	60	3.750	18.31	.0938	2.381	.0919	2.335
14	50	3.125	15.26	.0781	1.984	.0766	1.946
15	45	2.812	13.73	.0703	1.786	.0689	1.751
16	40	2.500	12.21	.0625	1.588	.0613	1.557
17	36	2.250	10.99	.0562	1.429	.0551	1.400
18	32	2.000	9.765	.0500	1.270	.0490	1.245
19	28	1.750	8.544	.0438	1.111	.0429	1.090
20	24	1.500	7.324	.0375	.952	.0368	.934
21	22	1.375	6.713	.0344	.873	.0337	.856
22	20	1.250	6.103	.0312	.794	.0306	.778
23	18	1.125	5.493	.0281	.714	.0276	.700
24	16	1.000	4.882	.0250	.635	.0245	.623
25	14	.8750	4.272	.0219	.556	.0214	.545
26	12	.7500	3.662	.0188	.476	.0184	.467
27	11	.6875	3.357	.0172	.437	.0169	.428
28	10	.6250	3.052	.0156	.397	.0153	.389
29	9	.5625	2.746	.0141	.357	.0138	.350
30	8	.5000	2.441	.0125	.318	.0123	.311
31	7	.4375	2.136	.0109	.278	.0107	.272
32	6 1/2	.4062	1.983	.0102	.258	.0100	.253
33	6	.3750	1.831	.0094	.238	.0092	.233
34	5 1/2	.3438	1.678	.0086	.218	.0084	.214
35	5	.3125	1.526	.0078	.198	.0077	.195
36	4 1/2	.2812	1.373	.0070	.179	.0069	.175
37	4 1/4	.2656	1.297	.0066	.169	.0065	.165
38	4	.2500	1.221	.0062	.159	.0061	.156
39	3 3/4	.2344	1.144	.0059	.149	.0057	.146
40	3 1/2	.2188	1.068	.0055	.139	.0054	.136
41	3 3/8	.2109	1.030	.0053	.134	.0052	.131
42	3 1/4	.2031	.9917	.0051	.129	.0050	.126
43	3 1/8	.1953	.9536	.0049	.124	.0048	.122
44	3	.1875	.9155	.0047	.119	.0046	.117

* For the Galvanized Sheet Gage, add 2.5 ounces to the weight per square foot as given in the table. Gage numbers below 8 and above 34 are not used in the Galvanized Sheet Gage.

† Gage numbers greater than 38 were not in the standard as set up by law, but are in general use.

Table 37. American Wire Gage—Weights of Copper, Aluminum, and Brass Sheets and Plates

Gage No.	Thickness		Approximate weight * per sq ft in lb		
	Inch	mm	Copper	Aluminum	Commercial (high) brass
0000	0.4600	11.68	21.27	6.49	20.27
000	.4096	10.40	18.94	5.78	18.05
00	.3648	9.266	16.87	5.14	16.07
0	.3249	8.252	15.03	4.58	14.32
1	.2893	7.348	13.38	4.08	12.75
2	.2576	6.544	11.91	3.632	11.35
3	.2294	5.827	10.61	3.234	10.11
4	.2043	5.189	9.45	2.880	9.00
5	.1819	4.621	8.41	2.565	8.01
6	.1620	4.115	7.49	2.284	7.14
7	.1443	3.665	6.67	2.034	6.36
8	.1285	3.264	5.94	1.812	5.66
9	.1144	2.906	5.29	1.613	5.04
10	.1019	2.588	4.713	1.437	4.490
11	.0907	2.305	4.195	1.279	3.996
12	.0808	2.053	3.737	1.139	3.560
13	.0720	1.828	3.330	1.015	3.172
14	.0641	1.628	2.965	0.904	2.824
15	.0571	1.450	2.641	.805	2.516
16	.0508	1.291	2.349	.716	2.238
17	.0453	1.150	2.095	.639	1.996
18	.0403	1.024	1.864	.568	1.776
19	.0359	0.9116	1.660	.506	1.582
20	.0320	.8118	1.480	.451	1.410
21	.0285	.7230	1.318	.402	1.256
22	.0253	.6438	1.170	.3567	1.115
23	.0226	.5733	1.045	.3186	0.996
24	.0201	.5106	0.930	.2834	.886
25	.0179	.4547	.828	.2524	.789
26	.0159	.4049	.735	.2242	.701
27	.0142	.3606	.657	.2002	.626
28	.0126	.3211	.583	.1776	.555
29	.0113	.2859	.523	.1593	.498
30	.0100	.2546	.4625	.1410	.4406
31	.00893	.2268	.4130	.1259	.3935
32	.00795	.2019	.3677	.1121	.3503
33	.00708	.1798	.3274	.0998	.3119
34	.00630	.1601	.2914	.0888	.2776
35	.00561	.1426	.2595	.0791	.2472
36	.00500	.1270	.2312	.0705	.2203
37	.00445	.1131	.2058	.0627	.1961
38	.00397	.1007	.1836	.0560	.1749
39	.00353	.0897	.1633	.0498	.1555
40	.00314	.0799	.1452	.0443	.1383

* Assumed specific gravities or densities in grams per cubic centimeter; copper, 8.89; aluminum, 2.71; brass, 8.47.

A No. 10 A.W.G. copper wire has the following approximate characteristics:

Ohms per 1000 ft.	1
Circular mils area.	10,000
Weight, pounds per 1000 ft.	32

A No. 10 A.W.G. aluminum wire has the following approximate characteristics:

Ohms per 1000 ft.	1.6
Circular mils area.	10,000
Weight, pounds per 1000 ft.	9.5

Remembering these rules it is easy to find the approximate size, resistance, area, or weight of any size wire. For example, a No. 12 A.W.G. copper wire has a resistance of 1 plus 60 per cent = 1.6 ohms per 1000 ft approximately. Its area, being inversely as its resistance, is 10,000/1.6 = 6250 circular mils; its diameter is therefore $\sqrt{6250} = 79$ mils, and its weight 32/1.6 = 20 lb per 1000 ft.

U. S. Steel Wire Gage. This gage, known also as the "Washburn and Moen," "Roebling," "American Steel and Wire Co.'s gage," is the one usually employed in the United States for steel and iron wire. It is frequently abbreviated "S.W.G.," but to avoid confusion with the British Standard Wire Gage (*see below*) it should be abbreviated "Stl. W.G." or "A. (steel) W.G."

Birmingham (or Stubs' Iron) Wire Gage. This gage is still used in the United States for some purposes, e.g., to designate the size of brass wire, and is also employed to a limited extent in Great Britain. It is usually abbreviated "B.W.G." It is sometimes referred to as the "Stubs' Iron Wire Gage," but it should not be confused with the Stubs' Steel Wire Gage.

British Standard Wire Gage. This gage, usually called simply the "Standard Wire Gage" and abbreviated "S.W.G.," is also known as the "New British Standard" (abbreviated "N.B.S."), the English Legal Standard, or the Imperial Wire Gage, and is the legal standard of Great Britain for all wires, as fixed by order in Council, August 23, 1883. It was constructed by modifying the Birmingham Wire Gage, so that the differences between successive diameters were the same for short ranges, i.e., so that a graph representing the diameters consists of a series of a few straight lines.

Edison Wire Gage. The size of a wire on this gage is equal to its cross-sectional area in circular mils divided by 1000. For example, a solid wire 0.2 in. in diameter has the number $(200)^2/1000 = 40$. This gage is now rarely used.

Metric Wire Gage. The gage number is ten times the diameter in millimeters.

Other Gages. In addition wire sizes are sometimes specified in terms of the "Old English Wire Gage," known also as the "London Gage," and the "Stubs' Steel Wire Gage." The Old English Wire Gage is the same as B.W.G. for all gage numbers under 20.

Comparison of Wire Gages. A comparison of the different gages, in terms of the diameters (in mils or thousandths of an inch) of solid wires corresponding to the various numbers, is given in Table 38. The cross-section in **circular mils** is the square of the diameter in mils.

Table 38. Comparison of Wire Gage Diameters in Mils
(Bureau of Standards, *Circulars* 31 and 67)

Gage No.	American wire gage (B. & S.)	Steel wire gage	Birmingham wire gage (Stubs')	Old English wire gage (London)	Stubs' steel wire gage	(British) Standard wire gage	Metric gage *	Gage No.
7–0	....	490.0	....	....	....	500	....	7–0
6–0	....	461.5	....	....	....	464	....	6–0
5–0	....	430.5	....	....	....	432	....	5–0
4–0	460	393.8	454	454	....	400	....	4–0
3–0	410	362.5	425	425	....	372	....	3–0
2–0	365	331.0	380	380	....	348	....	2–0
0	325	306.5	340	340	....	324	....	0
1	289	283.0	300	300	227	300	3.94	1
2	258	262.5	284	284	219	276	7.87	2
3	229	243.7	259	259	212	252	11.8	3
4	204	225.3	238	238	207	232	15.7	4
5	182	207.0	220	220	204	212	19.7	5
6	162	192.0	203	203	201	192	23.6	6
7	144	177.0	180	180	199	176	27.6	7
8	128	162.0	165	165	197	160	31.5	8
9	114	148.3	148	148	194	144	35.4	9
10	102	135.0	134	134	191	128	39.4	10
11	91	120.5	120	120	188	116	....	11
12	81	105.5	109	109	185	104	47.2	12
13	72	91.5	95	95	182	92	....	13
14	64	80.0	83	83	180	80	55.1	14
15	57	72.0	72	72	178	72	....	15
16	51	62.5	65	65	175	64	63.0	16
17	45	54.0	58	58	172	56	....	17
18	40	47.5	49	49	168	48	70.9	18
19	36	41.0	42	40	164	40	....	19
20	32	34.8	35	35	161	36	78.7	20
21	28.5	31.7	32	31.5	157	32	....	21
22	25.3	28.6	28	29.5	155	28	....	22
23	22.6	25.8	25	27.0	153	24	....	23
24	20.1	23.0	22	25.0	151	22	....	24
25	17.9	20.4	20	23.0	148	20	98.4	25
26	15.9	18.1	18	20.5	146	18	....	26
27	14.2	17.3	16	18.75	143	16.4	....	27
28	12.6	16.2	14	16.50	139	14.8	....	28
29	11.3	15.0	13	15.50	134	13.6	....	29
30	10.0	14.0	12	13.75	127	12.4	118	30
31	8.9	13.2	10	12.25	120	11.6	....	31
32	8.0	12.8	9	11.25	115	10.8	....	32
33	7.1	11.8	8	10.25	112	10.0	....	33
34	6.3	10.4	7	9.50	110	9.2	....	34
35	5.6	9.5	5	9.00	108	8.4	138	35
36	5.0	9.0	4	7.50	106	7.6	....	36
37	4.5	8.5	....	6.50	103	6.8	....	37
38	4.0	8.0	....	5.75	101	6.0	....	38
39	3.5	7.5	....	5.00	99	5.2	....	39
40	3.1	7.0	....	4.50	97	4.8	157	40
41	....	6.6	....	....	95	4.4	....	41
42	....	6.2	....	....	92	4.0	....	42
43	....	6.0	....	....	88	3.6	....	43
44	....	5.8	....	....	85	3.2	....	44
45	....	5.5	....	....	81	2.8	177	45
46	....	5.2	....	....	79	2.4	....	46
47	....	5.0	....	....	77	2.0	....	47
48	....	4.8	....	....	75	1.6	....	48
49	....	4.6	....	....	72	1.2	....	49
50	....	4.4	....	....	69	1.0	197	50

* For diameters corresponding to metric gage numbers, 1.2, 1.4, 1.6, 1.8, 2.5, 3.5, and 4.5, divide those of 12, 14, etc., by ten.

SYMBOLS AND ABBREVIATIONS

19. ABBREVIATIONS FOR ENGINEERING TERMS

NOTE: This list is a selection of American Tentative Standard abbreviations, for scientific and engineering terms, recommended by the American Standards Association. (See ASA, Z10.1—1941.)

Absolute.............................. abs
Acre..............................spell out
Alternating-current (as adjective).......... ac
Ampere.............................. amp
Ampere-hour......................... amp-hr
Angstrom unit.......................... A
Atmosphere........................... atm
Atomic weight...................... at. wt.
Average.............................. avg
Avoirdupois.......................... avdp

Barometer............................ bar.
Barrel................................ bbl
Baumé................................ Bé
Boiler pressure..................... spell out
Boiling point.......................... bp
Brake horsepower...................... bhp
Brake horsepower-hour................ bhp-hr
Brinell hardness number................. Bhn
British thermal unit.......... .... Btu or B

Calorie............... .... cal
Candle.................................. c
Candle-hour.......................... c-hr
Candlepower........................... cp
Centigram............................. cg
Centiliter.............................. cl
Centimeter............................ cm
Centimeter-gram-second (system).......... cgs
Chemically pure.......................... cp
Circular............................... cir
Circular mils....................... cir mils
Coefficient........................... coef
Cologarithm.......................... colog
Concentrate.......................... conc
Conductivity.......................... cond
Constant............................. const
Cord................................... cd
Cosecant.............................. csc
Cosine................................ cos
Cotangent............................. cot
Coulomb.......................... spell out
Counter electromotive force............. cemf
Cubic.................................. cu
Cubic centimeter............. cu cm, cm^3, cc
Cubic feet per minute................... cfm
Cubic foot............................cu ft
Cubic inch.......................... cu in.
Cubic meter.................. cu m or m^3
Cubic yard.......................... cu yd
Cycles per second............... spell out or c

Decibel................................ db
Degree.......................... deg or °
Degree Centigrade....................... C
Degree Fahrenheit....................... F
Degree Kelvin........................... K
Degree Réaumur......................... R
Diameter............................. diam
Direct-current (as adjective)............. d-c
Dozen................................ doz
Dram.................................. dr

Efficiency............................. eff
Electric.............................. elec
Electromotive force..................... emf
Equation................................ eq
External............................... ext

Farad........................ spell out or f
Foot..................................... ft
Foot-candle............................ ft-c
Foot-Lambert.......................... ft-L
Foot-pound........................... ft-lb
Foot-pound-second (system)............... fps
Freezing point........................... fp
Fusion point........................... fnp

Gallon................................ gal
Grain............................ spell out
Gram.................................... g
Gram-calorie .... g-cal

Henry.................................... h
Horsepower............................. hp
Horsepower-hour...................... hp-hr
Hour.................................... hr
Hundred................................. C
Hyperbolic sine........................ sinh
Hyperbolic cosine...................... cosh
Hyperbolic tangent..................... tanh

Inch.................................... in.
Inch-pound........................... in-lb
Internal................................ int

Joule..................................... j

Kilocycles per second.................... kc
Kilogram................................ kg
Kilogram-calorie...................... kg-cal
Kilogram-meter........................ kg-m
Kiloliter................................. kl
Kilometer............................... km
Kilovolt.................................. kv
Kilovolt-ampere......................... kva
Kilowatt................................. kw
Kilowatthour.......................... kwhr

Lambert.................................. L
Latitude........................... lat or ϕ
Linear foot............................ lin ft
Liter...................................... l
Liquid.................................... liq
Logarithm (common)...................... log
Logarithm (natural) .... $\log_e$ or ln
Longitude......................... long. or λ
Lumen..................................... l
Lumen-hour............................. l-hr

Magnetomotive force...................... mmf
Maximum .... max
Melting point............................ mp
Meter..................................... m
Meter-kilogram......................... m-kg
Mho.............................. spell out
Microampere...................... μa or mu a
Microfarad.............................. μf
Micromicron................... $\mu\mu$ or mu mu
Micron............................. μ or mu
Microwatt....................... μw or mu w
Mile.............................. spell out
Milliampere.............................. ma
Milligram................................ mg
Millihenry............................... mh

Abbreviations for Engineering Terms—*Continued*

Milliliter	ml
Millimeter	mm
Millimicron	mμ or m mu
Million	spell out
Millivolt	mv
Mean horizontal candlepower	mhcp
Miles per hour	mph
Minimum	min
Minute	min
Minute (angular measure)	′
Ohm	spell out or Ω
Ounce	oz
Ounce-foot	oz-ft
Ounce-inch	oz-in.
Pint	pt
Potential	spell out
Pound	lb
Pound-foot	lb-ft
Pound-inch	lb-in.
Pounds per square foot	psf
Pounds per square inch	psi
Power factor	spell out or pf
Quart	qt
Radian	spell out
Reactive kilovolt-ampere	kvar
Reactive volt-ampere	var
Revolutions per minute	rpm
Revolutions per second	rps
Root mean square	rms
Secant	sec
Second	sec
Second (angular measure)	″
Sine	sin
Specific gravity	sp gr
Specific heat	sp ht
Spherical candlepower	scp
Square	sq
Square centimeter	sq cm or cm^2
Square foot	sq ft
Square inch	sq in.
Square kilometer	sq km or km^2
Square meter	sq m or m^2
Square micron	sq μ or sq mu or μ^2
Square root of mean square	rms
Standard	std
Tangent	tan
Temperature	temp
Thousand	M
Ton	spell out
Versed sine	vers
Volt	v
Volt-ampere	va
Watt	w
Watthour	whr
Weight	wt
Yard	yd
Year	yr

20. LETTER SYMBOLS FOR THE MAGNITUDES OF ELECTRICAL QUANTITIES

(Tentative American Standard Z10.5–1947) †

In the alphabetical order of the names of the quantities

Each quantity appears at only one place in this table (with a few exceptions), listed alphabetically under its preferred name. The non-preferred names appear in parentheses under the preferred names. Deprecated names are also in parentheses and in addition are asterisked thus: (electric force)*.

Names beginning with the qualifying adjectives, *electric*, *electrostatic*, *dielectric*, *magnetic*, *mutual*, *self*, and *relative*, are listed under the term that is so qualified.

Symbols for scalar quantities, whose values are expressed by real numbers, are printed in ordinary-face *italic letters.*

Symbols for vector quantities are printed in **bold-face Roman letters.**

Symbols for phasor quantities, whose values are expressed by complex numbers, are printed in ***bold-face italic letters.***

Item	Quantity	Symbol
1	admittance	Y
2	attenuation constant	α
3	capacitance	C
	(capacity) *	
	(permittance) *	
4	capacitivity	ϵ
	dielectric constant	
	(permittivity) *	
	of evacuated space	ϵ_v
5	capacitivity, relative	ϵ_r
	relative dielectric constant	
	(specific inductive capacity) *	
6	charge, electric or quantity of electricity	Q
	charge density	
7	line d. of charge	λ
8	surface d. of charge	σ
9	volume d. of charge	ρ
10	conductance	G
11	conductivity	γ
12	conductivity, equivalent	Λ
13	coupling coefficient	k
14	current	I
	(intensity of current) *	
15	current density	
16	sheet c.d. (linear c.d.)	A
17	damping constant or coefficient	δ
	(decay constant)	

* Deprecated name.

† Reprinted by permission of the American Institute of Electrical Engineers.

Letter Symbols for the Magnitudes of Electrical Quantities—*Continued*

Item	Quantity	Symbol
18	dielectric constant see capacitivity	ϵ
	dielectric, a qualifier see term that it qualifies	
19	displacement, electric	$\mathbf{D}$
20	efficiency	η
21	elastance mutual e. S_m, S_{rc} self e. S, S_{cc}	S
22	elastivity	σ
	electric, a qualifier see term that it qualifies	
23	electronic charge (absolute value of)	e
24	electromotive force (electromotance) (potential difference, electric) (voltage) *	E
25	energy	W
26	force	$\mathbf{F}$
27	flux, displacement f. (flux of e. displacement)	Ψ
28	flux, magnetic (flux of magnetic induction)	Φ
29	flux-linkage	
30	frequency	f
31	frequency, angular angular velocity	ω
32	frequency, rotational	n
33	impedance mutual i. Z_m, Z_{rc} self i. Z, Z_{cc}	Z
34	induction, magnetic (magnetic flux density)	$\mathbf{B}$
35	inductance mutual i. L_m, L_{rc} self i. L, L_{cc}	L
36	intensity, electric (electric field intensity) (electric field strength) (electric force) * (electric field) *	$\mathbf{E}, \mathbf{K}$
37	intensity, magnetic or magnetizing force (magnetic field strength) (magnetic force) *	$\mathbf{H}$
	magnetic, a qualifier see term that it qualifies	
38	magnetomotive force (m. potential difference) magnetomotance	$M, \mathfrak{F}$
39	moment, electric	$\mathbf{p}$
40	moment, magnetic	$\mathbf{m}$
41	number of conductors or turns	N
42	number of poles	p
43	number of phases	m
44	period	T
45	permeance	$\mathcal{P}, \Lambda$
46	permeability, magnetic of evacuated space	μ μ_v
47	permeability, relative	μ_r
48	(permittivity) * (see capacitivity)	
49	phase angle	φ
50	phase constant wavelength constant (wave number)	β
51	polarization, electric	
52	polarization, magnetic intrinsic induction metallic induction	$\mathbf{B_i}$
53	pole strength	m
54	potential, electric (electromagnetic scalar p.)	V
55	potential, retarded scalar	
56	potential, magnetic (magnetic scalar p.) m. pot. difference	$M, \mathfrak{F}$
57	potential, magnetic vector p.	$\mathbf{A}$
58	potential, retarded vector p.	$\mathbf{A_r}$
59	power, active	P
60	power, reactive volt-amperes, reactive	Q
61	power, apparent volt-amperes	S
62	power factor	F_p
63	propagation constant	γ
64	Poynting vector	$\mathbf{\Pi}$
65	quantity of electricity charge, electric	Q
66	quality factor of a reactor figure of merit of a reactor	Q
67	reactance capacitative r. inductive r. mutual r. X_m, X_{rc} self r. X, X_{cc}	X X_c X_L
68	reactive factor	F_q
69	reluctance	$\mathcal{R}$
70	reluctivity	ν
71	resistance mutual r. R_m, R_{rc} self r. R, R_{cc}	R
72	resistivity	ρ
73	resistance-temperature coefficient	α
	rotative operators	
74	90°, $\sqrt{-1}$	j
75	120°, $\sqrt[3]{1}$	a
	self, a qualifier see term that it qualifies	
76	slip	s
77	susceptance	B
	susceptibility	
78	dielectric s. intrinsic capacitivity	η
79	magnetic s. intrinsic permeability	κ
80	symmetrical components (Note 5)	
81	temperature	t, (θ) T, (Θ)
82	time	t
83	time constant	τ
84	velocity of light	c
85	vibration constant (oscillation constant)	$\mathbf{p}$
86	wavelength	λ
87	wavelength constant phase constant	β
88	work	W

* Deprecated name.

Note 1. Designation of maximum, instantaneous, rms, and average values.

Where distinctions between maximum, instantaneous, root-mean-square (effective), and average values are necessary, E_m, I_m, Q_m, and P_m are recommended for maximum values; e, i, q, and p for instantaneous values, E, I, and Q for root-mean-square values and E_a, I_a, Q_a, and P for average values.

Note 2. Quantities per unit volume, area, or length.

It is recommended that quantities per unit volume, area, length, etc., be represented as far as practicable by lower-case letters corresponding to the cap letters which represent the total quantities, or by the cap letters with the subscript 1, except for those quantities for which this table has symbols for the quantity per unit volume, area, etc.

Note 3. Distinction between the symbols V and E for potential and electromotive force.

The distinction between the use of V for potential and E for electromotive force is:

V is to be used for potentials or potential differences that are attributed solely to that distribution of electric field intensities which is computed (by the inverse square law of force) from the segregated charges of the field.

E is to be used for the emf along a path from a terminal A to a terminal B when in the region A to B one or more non-electrostatic types of electric intensities exist. or turbulent actions occur—as in voltaic cells, electrostatic generators, and electromagnetic sources of emf.

Note 4. The sequence of the double subscripts to multiplying operators.

The sequence of the double subscripts to the *multiplying* operators (mutual impedances, resistances, or elastances or transconductances, etc.) that occur in the fundamental equations of networks is to be determined by the following consideration:

Consideration. The set of fundamental equations (e.g., Kirchhoff's emf equations) should yield a determinant in which the subscript sequence conforms to the mathematician's convention for writing determinants; namely,

Convention. In the double subscripts of the elements of a determinant, the subnumber designating the "row" is to precede the subnumber designating the "column" to which the element belongs, or the order is e_{rc}. Thus

$$D(e) = \begin{vmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{vmatrix}$$

This consideration leads to the following rule:

Rule for writing double subscripts.

The first subnumber in the symbol for a *multiplying* operator designates the number of the circuit in which the *product* of the *multiplication* is measurable, while the second subnumber designates the number of the circuit in which the operand or *multiplicand* is measurable.

As an illustration, Kirchhoff's emf law for the emfs of the rth circuit due to the currents in all the circuits of a network is written:

$$E_{r,d} = Z_{r1}I_1 + Z_{r2}I_2 + \cdots Z_{rr}I_r + \cdots$$

or

$$E_{r,d} = \Sigma c\ Z_{rc}I_c$$

($E_{r,d}$ being the driving emf impressed in the rth circuit).

Note 5. Notation for symmetrical components.

The standard notation for designating the symmetrical components of the currents and potential differences in unbalanced polyphase systems is that subscript notation in which:

(a) double subscripts are added to the symbols for current and potential difference;

(b) the first and second subscripts designate, respectively, the phase and the sequence to which the component belongs;

(c) the first, or phase, subscript may be the phase number, or the phase letter, or a two-letter combination that designates (on a diagram) both the phase and the direction in the phase;

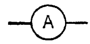
(d) the second, or sequence, subscript is *always* to be the number that designates the sequence to which the component belongs; the positive, negative, and zero sequence components in three-phase systems being designated by the numbers 1, 2, and 0, respectively.

Illustration of notation:

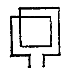
$$I_a = I_{a1} + I_{a2} + I_{a0}$$
$$I_b = I_{b1} + I_{b2} + I_{b0}$$
$$I_c = I_{c1} + I_{c2} + I_{c0}$$

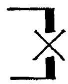
21. STANDARD GRAPHICAL SYMBOLS

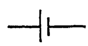
(Approved by American Standards Association, Nov. 1, 1942)
(Revised by American War Standard, April 18, 1944)

1. Ammeter 

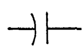
2. Antenna 

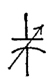
3. Antenna, Loop 

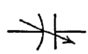
4. Arc 

5. Battery
   Long line always positive but polarity may be indicated in addition. 

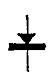
6. Capacitor, Fixed
   Condenser, Fixed
   The curved electrode identifies the outermost electrode where applicable or the negative electrode for electrolytic capacitors. 

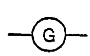
7. Capacitor, Fixed, Shielded 

8. Capacitor, Variable
   The curved portion is the movable electrode. 

9. Capacitor, Variable, Shielded 

10. Counterpoise 

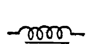
11. Crystal Detector 

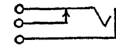
12. Galvanometer 

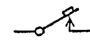
13. Ground 

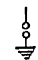
14. Inductor

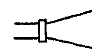
15. Inductor, Adjustable or Variable  

16. Inductor Iron Core 

17. Jack 

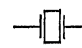
18. Key 

19. Lightning Arrester 

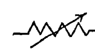
20. Loudspeaker 

21. Microphone 

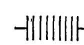
22. Phototube 

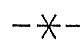
23. Piezoelectric Plate 

24. Resistor 

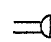
25. Resistor, Adjustable or Variable 

26. Spark Gap, Plain 

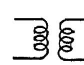
27. Spark Gap, Quenched 

28. Spark Gap, Rotary 

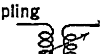
29. Telephone Receiver 

30. Telephone Transmitter 

31. Thermoelement 

32. Transformer, Air Core 

33. Transformer, Iron Core 

34. Transformer, with Variable Coupling 

35. Voltmeter 

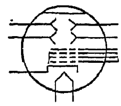
36. Wires, Crossed, not joined

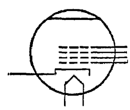
37. Wires, Joined 

Electron Tubes

(ASA Z32.10–1944)

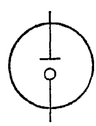
1. Anode or Plate (Including Collector)


2. Cathode-Ray Tube with Electrostatic Deflection


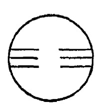
3. Cathode-Ray Tube for Magnetic Deflection


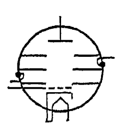
4. Cold Cathode (Including Ionic-Heated Cathode)


5. Deflecting, Reflecting, or Repelling Electrode (Electrostatic Type)

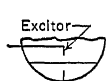

6. Diode (Cold-cathode and Gas Content)


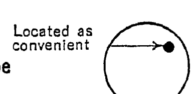
7. Directly Heated Cathode (Filament Type)


8. Double-Cavity Resonator Envelope


9. Double-Cavity Velocity-Modulation Tube with Collecting Electrode


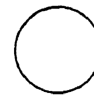
10. Dynode

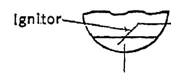

11. Excitor (Contactor Type)


12. Gas-Filled Envelope


13. Grid (Including Beam-Confining or Beam-Forming Electrodes)

14. Heater

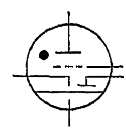

15. High-Vacuum Envelope


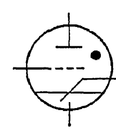
16. Ignitor


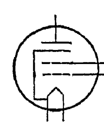
17. Indirectly Heated Cathode


18. Ionic-Heated Cathode with Supplementary Heater

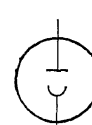

19. Loop Coupling (Electromagnetic Type)

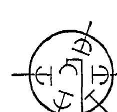

20. Mercury Pool Tube with Excitor, Control Grid, and Holding Anode


21. Mercury Pool Tube with Ignitor and Control Grid


22. Pentode (Suppressor or Beam-confining Electrodes)


23. Photoelectric Cathode

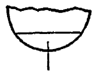

24. Phototube


25. Phototube (Multiplier Type)


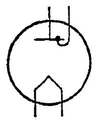
Electron Tubes—*Continued*

(ASA Z32.10–1944)

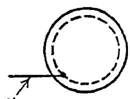
26. Pool Cathode



27. Resonant Magnetron

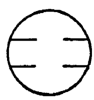


28. Shield within Envelope

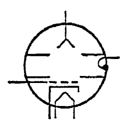


External Connection

29. Single-Cavity Resonator Envelope

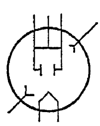


30. Single-Cavity Velocity-Modulation Tube with Reflecting Electrode

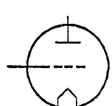


31. Target, X-Ray

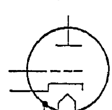
32. Transit-Time Split-Plate Type Magnetron with Stabilizing Deflecting Electrodes and Internal Circuit



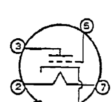
33. Triode with Filamentary Cathode



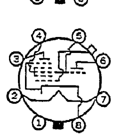
34. Triode with Indirectly Heated Cathode and Envelope Connection



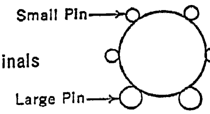
35. Triode with Indirectly Heated Cathode and Envelope Connected to Base Terminal



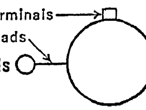
36. Triode-Heptode with Rigid Envelope Connection



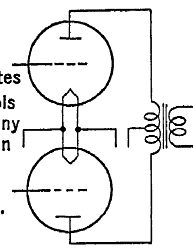
37. Tube Base Terminals



38. Tube Envelope Terminals



39. This figure illustrates how tube symbols may be placed in any convenient position as shown in a communication transformer circuit.



40. General Notes

(*a*) The diagram for a tube having more than one heater shall show only one heater symbol (inverted V) unless the heaters have entirely separate connections. If a tap is made, one heater symbol shall still be shown, and the tap shall be shown at the vertex of the heater symbol, regardless of the actual division of voltage across the heater

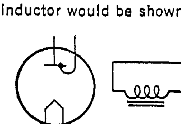
(*b*) Item (*a*) shall apply also to filaments. In case of a tap, either brought out to a pin connection or internally connected as to a suppressor grid, the tap shall be shown at the vertex of the filament symbol, regardless of the actual division of voltage across the filament.

(*c*) A type having more than one cathode shall be shown as having a single cathode unless separate cathode connections are made.

(*d*) A type having two or more grids tied internally shall be shown with symbols for each grid, except when the grids are adjacent in the tube structure. Thus the diagram for a twin pentode having a common screen-grid connection for each section and for a converter tube having the No. 3 and the No. 5 grids connected internally will show separate symbols for each grid. However, a triode where the control grid is physically in the form of two grid windings would show only one grid.

(*e*) A type having a grid adjacent to a plate but internally connected to the plate to form a portion of it shall be shown as having a plate only,

(*f*) Associated parts of a circuit such as deflecting coils, field coils, etc., are not a part of the tube symbol but may be added to the circuit in the form of standard symbols as shown in ASA Z32.3 or ASA Z32.5. For example, resonant-type magnetron plus symbol for ferromagnetic inductor would be shown



22. USE OF GREEK ALPHABET FOR SYMBOLS

Capital	Lower Case	Name	Commonly Used to Designate
Α	α	Alpha	Angles. Area. Coefficients. Attenuation constant.
Β	β	Beta	Angles. Flux density. Coefficients.
Γ	γ	Gamma	Conductivity. Specific gravity. Propagation constant.
Δ	δ	Delta	Variation. Density. Damping coefficient.
Ε	ϵ	Epsilon	Base of natural logarithms. Capacitivity.
Ζ	ζ	Zeta	Impedance. Coefficients. Coordinates.
Η	η	Eta	Hysteresis coefficient. Efficiency.
Θ	θ	Theta	Temperature. Phase angle.
Ι	ι	Iota	
Κ	κ	Kappa	Dielectric constant. Susceptibility.
Λ	λ	Lambda	Wavelength.
Μ	μ	Mu	Micro. Amplification factor. Permeability.
Ν	ν	Nu	Reluctivity.
Ξ	ξ	Xi	
Ο	o	Omicron	
Π	π	Pi	Ratio of circumference to diameter = 3.1416.
Ρ	ρ	Rho	Resistivity.
Σ	σ	Sigma	Capital: sign of summation.
Τ	τ	Tau	Time constant. Time phase displacement.
Υ	υ	Upsilon	
Φ	ϕ or φ	Phi	Magnetic flux. Angles.
Χ	χ	Chi	
Ψ	ψ	Psi	Dielectric flux. Phase difference.
Ω	ω	Omega	Capital: ohms. Lower case: angular velocity, or $2\pi \times$ frequency.

CONSTANTS

By Carl C. Chambers

23. PRINCIPAL PHYSICAL CONSTANTS AND RATIOS *

Velocity of light	$(2.99776 \pm 0.00004) \times 10^{10}$ cm sec^{-1}
Ratio of electrostatic to electromagnetic units	$(2.9971 \pm 0.0001) \times 10^{10}$ cm$^{1/2}$ sec$^{-1/2}$ (int ohms)$^{1/2}$ $(2.9978 \pm 0.0001) \times 10^{10}$ cm sec^{-1} (in absolute units)
Volume of a perfect gas (0 deg cent and normal atmospheric pressure)	$(22.4146 \pm 0.0006) \times 10^3$ cm^3 mole^{-1}
Normal atmospheric pressure	$(1.013246 \pm 0.000004) \times 10^6$ dynes cm^{-2}
45 deg cent atmospheric pressure	$(1.013195 \pm 0.000004) \times 10^6$ dynes cm^{-2}
Ice point (absolute scale)	$273.18 \pm 0.01°$ K
Mechanical equivalent of heat (15 deg cent)	4.1855 ± 0.0004 abs joule cal^{-1}
Electrical equivalent of heat (15 deg cent)	4.1847 ± 0.0003 int joule cal^{-1}
Faraday constant	96494 ± 5 int coulombs g-equiv^{-1}
Electronic charge	$(4.8025 \pm 0.0010) \times 10^{-10}$ abs-es unit $(1.60203 \pm 0.00034) \times 10^{-20}$ abs-em unit
Planck constant	$(6.624 \pm 0.002) \times 10^{-27}$ erg sec
Acceleration of gravity	980.665 cm sec^{-2}
Electrochemical equivalent of silver	1.11800×10^{-3} g. int coulombs^{-1}
Wave length of red cadmium line (15 deg cent, normal atmospheric pressure)	6438.4696 I.A. †
Effective grating space of calcite (18 deg cent)	3.02904×10^{-8} cm
Avogadro's number	$(6.0228 \pm 0.0011) \times 10^{23}$ mole^{-1}
Boltzmann constant	$(1.3708 \pm 0.0014) \times 10^{-16}$ erg deg^{-1}
Stefan-Boltzmann constant	$(5.672 \pm 0.003) \times 10^{-5}$ erg cm^{-2} deg^{-4} sec^{-1}
Mass of the electron	$(9.1066 \pm 0.0032) \times 10^{-28}$ g
Ratio of mass of H to mass of electron (measured by deflection)	1837.5 ± 0.5

* Values taken from Birge, *Rev. of Mod. Phys.*, Vol. 13, No. 4 (October, 1941).

† This defines the international angstrom unit (I.A.). The unit is of the order of 1 part in several million different from 10^{-8} cm.

24. STANDARD RADIO-FREQUENCY RANGES

The International Telecommunication Union in 1947 adopted officially the Nomenclature of Frequencies shown in Table 39. The designation of each metric subdivision of wavelength range is the name of the metric unit of length which is equal to the shortest wavelength in the range. The range number is the power of 10 which represents the approximate mean frequency of that range. The frequency subdivision designations are those first adopted by the Armed Forces of the United States and subsequently by other branches of the government.

The table is based on the approximation that the wave velocity is 300,000,000 meters per second. In any case, where the required precision makes this assumption inadequate, the exact boundaries of the ranges should be based on the frequency range, not the wavelength range.

It is suggested that the term radio be added where there is any possibility of confusion. A power engineer might be considerably startled to hear 100 kc referred to as a "low frequency."

The term microwave has been variously used: (1) as referring to waves less than 1 meter in length; (2) as certainly including ranges 10 and 11, and part of 9; (3) as referring to waves using cavities instead of *LC* tuned circuits; and (4) as referring to waves transmitted by wave guides. Where any confusion is possible, the term should be eschewed or clarifying text added.

Table 39. Nomenclature of Frequencies

Range Number N	Frequency Range		Wavelength Range		Metric Subdivision	Frequency Subdivision
	Lower Limit (Incl.)	Upper Limit (Excl.)	Lower Limit (Excl.)	Upper Limit (Incl.)		
0	0.3 c	3 c				
1	3 c	30 c				
2	30 c	300 c				
3	300 c	3 kc				
4	3 kc	30 kc	10 km	100 km	Myriametric waves	VLF (very low frequency)
5	30 kc	300 kc	1 km	10 km	Kilometric waves	LF (low frequency)
6	300 kc	3,000 kc	1 hm	10 hm	Hectometric waves	MF (medium frequency)
7	3,000 kc	30,000 kc	1 dkm	10 dkm	Decametric waves	HF (high frequency)
8	30,000 kc	300 Mc	1 m	10 m	Metric waves	VHF (very high frequency)
9	300 Mc	3,000 Mc	1 dm	10 dm	Decimetric waves	UHF (ultra high frequency)
10	3,000 Mc	30,000 Mc	1 cm	10 cm	Centimetric waves	SHF (super high frequency)
11	30,000 Mc	300 kMc	1 mm	10 mm	Millimetric waves	EHF (extremely high frequency)
12	300 kMc	3,000 kMc				
13	3,000 kMc	30,000 kMc				
14	30,000 kMc	300 MMc				

Note 1. Ranges 0–3 and 12–14 not standardized by I.T.U.

Note 2. Frequencies shall be expressed in kilocycles per second (kc/s) at and below 30,000 kilocycles per second and in megacycles per second (Mc/s) above this frequency.

Note 3. Used as an adjective the word "Range" shall precede the number; thus: "Range 3."

SECTION 2

PROPERTIES OF MATERIALS

CONDUCTING MATERIALS

By Knox McIlwain

INSULATING MATERIALS

By W. R. Dohan

MAGNETIC MATERIALS

By R. M. Bozorth and R. A. Chegwidden

PROPERTIES OF MATERIALS

CONDUCTING MATERIALS

By Knox McIlwain

Since all materials possess, to some extent, the ability to conduct electricity, whether a particular material is called conducting or insulating is a matter only of relative degree. In general, if a moderate potential difference, say from a voltaic cell, is placed across a section of a material and a measurable current flows the substance is said to be conducting, but if no conveniently detectable current flows it is considered insulating.

Among conducting materials are included pure metals, some metallic salts and oxides, alloys, and the metalloids, carbon, silicon, and boron. Such substances as glass, dry paper and silk, porcelain, and rubber possess such a very low conductivity that they are considered insulating materials.

1. DEFINITIONS

Conductivity. The conductivity of a material is the direct current conductance between the opposite, parallel faces of a portion of the material having unit length and unit cross-section.

Effective Conductivity. The effective conductivity of a material to a periodic current is the effective conductance between the opposite, parallel faces of a portion of the material having unit length and unit cross-section.

Resistivity. The resistivity of a material is the reciprocal of its conductivity.

Units of Resistivity. The resistivity may be expressed as the resistance of a 1-cm cube of material; this unit is called the *ohm per centimeter cube* or preferably simply the *ohm-centimeter*. In the English system the *ohm-inch* is used. When the material is to be drawn into wires the *ohm per mil-foot* is used; this is the resistance of a wire 1 mil (0.001 in.) in diameter and 1 ft long.

Units of Conductivity. The conductivity of a material γ is numerically equal to the reciprocal of its resistivity ρ and is expressed in mhos and megmhos per centimeter, etc., instead of ohm-centimeters and microhm-centimeters, etc., as used to express resistivity. Thus γ equals $1/\rho$.

Annealed Copper Standard. The standard, or 100 per cent, conductivity is defined as follows:

1. At a temperature of 20 deg cent, the resistance of a wire of standard annealed copper 1 meter in length and of a uniform section of 1 sq mm is $^1/_{58}$ ohm = 0.01724 ohm.

2. At a temperature of 20 deg cent, the density of standard annealed copper is 8.89 grams per cubic centimeter. This corresponds to 8.90 grams per cubic centimeter at 0 deg cent.

3. Thus, at 20 deg cent, the resistance of a wire of standard annealed copper of uniform section, 1 meter in length and weighing 1 gram, is $(^1/_{58}) \times 8.89 = 0.15328$ ohm.

Temperature Coefficient of Electric Resistance. The resistance temperature coefficient β_t of a substance at any temperature t is defined as the rate of change of resistance at this temperature divided by the resistance R_t at this temperature:

$$\beta_t = \frac{1}{R_t}\left(\frac{dR}{dt}\right)_t$$

The "mean" temperature coefficient α_t between any two temperatures t and t_1 "referred to" the temperature t is defined as the "average" change in the resistance in this interval per degree change of temperature, divided by the resistance at the lower temperature:

$$\alpha_t = \frac{R_{t1} - R_t}{R_t(t_1 - t)}$$

The temperature coefficients of 100 per cent conductivity copper are given in Table 1.

Table 1. Temperature Coefficients of Copper

Ohms per Meter-gram at 20° C	Per Cent Conductivity	α_0	α_{15}	α_{20}	α_{25}	α_{30}
0.16134	95	0.00403	0.00380	0.00373	0.00367	0.00360
.15966	96	.00408	.00385	.00377	.00370	.00364
.15802	97	.00413	.00389	.00381	.00374	.00367
.15753	97.3	.00414	.00390	.00382	.00375	.00368
.15640	98	.00417	.00393	.00385	.00378	.00371
.15482	99	.00422	.00397	.00389	.00382	.00374
<u>.15328</u>	<u>100</u>	.00427	.00401	<u>.00393</u>	.00385	.00378
.15176	101	.00431	.00405	.00397	.00389	.00382

The underlined values in the table have been adopted as standard by the American Institute of Electrical Engineers.

2. PROPERTIES OF SPECIFIC CONDUCTORS

Conducting materials can be roughly divided into two groups: the good conductors, and the resistive conductors. The good conductors are all *metals*, a group of some 50 chemical elements recognized as such by their hardness, ductility, malleability, luster, and good conductivity of heat and electricity. Since these and other characteristics are possessed by these elements in varying degrees there are some which are metallic in some properties and non-metallic in others. Of these an electrically important group is carbon, silicon, and boron, often called the metalloids, which are rather good conductors of heat and electricity but have non-metallic mechanical properties. Almost all the resistive conductors on the market today are solid solution alloys or those composed largely of solid solutions. In electric circuits they are generally used either in devices for purposes of operation, protection, or control, or as heating elements. It is usually desirable for them to have properties of high resistivity and low temperature coefficient of resistance, but in some cases a high temperature coefficient is useful—witness the use of nickel as a filament control.

In general, the standard alloys for electrical resistance are made of nickel and chromium, compositions of 80 per cent nickel and 20 per cent chromium having high resistance to oxidation with maximum working temperatures up to about 1100 deg cent. By varying the proportions of these and by the addition of different amounts of iron, copper, manganese, zinc, and cobalt we are able to obtain alloys which have different resistivities, temperature coefficients, melting points, magnetic properties, and heat- and corrosion-resisting properties.

Tables 2 and 3 give the properties of materials available for the manufacture of conductors and resistors; Table 4 gives the physical properties of some beryllium-copper alloys.

Table 2. Properties of Conducting Materials at Usual Temperatures

Description of Material	Resistivity Microhm-cm, 0° C	Resistivity Microhm-cm, 20° C	Temperature Coefficient, See Note a: Temp., °C	Temperature Coefficient, See Note a: α	Max. Working Temp., °C	Density, $\frac{gm}{cm^3}$	Tensile Str. (annealed), $\frac{lb}{in.^2}$	Coef. Lin. Expansion per °C $10^{-6}\times$
Acme * (Ni 30 + Cr 5 + Fe 65)		87.3	20	0.00072	1,000	8.15	100,000	
Advance * (Ni 45 + Cu 55)		48.8	20–100	.00002	535	8.9	60,000	14.9
Akbar * (Ni + Cu + Mn)		83.1	20	.0006	600		80,000	14
Alferon * (Cr 14.25 + Al 3.5 + Fe 82.25)		112	20–500	.00016	1,100	7.31	100,000	11.4
Alloy A	(Similar Chromel A)							
Alloy C	(Similar Chromel C)							
Alloy D	(Similar Comet)							
Alumel * (Ni 94 + Mn 2.5 + Fe 0.5 + Al 2 + Si 1)	33.3		0	.0012	1,250			
Aluminum (Pure)	2.62		0–100	.00423				
Aluminum (Wire, 61% cond.)	2.607	2.828	18	.0039	300	2.7	35,000	24
Aluminum Bronze (Cu 97 + Al 3)	8.85		15	.000897				
Antimony	39.1		20	.0036				
Argentan * (Cu 61.6 + Ni 15.8 + Zn 22.6)	28.5		0–160	.000387				
Argentan * (Cu 56 + Ni 26 + Zn 18)		(15°) 42	(Similar German Silver)					
Arsenic	35			.0042				
Ascoloy * (Fe 82–86 + Cr 16–12 + Mn 0.5 + Ni 0.5 + Si 0.5)					1,230			
Beryllium		10.1						
Bismuth		120	0–100 20	.00424 .004				
Boron		8×10^{12}						
Brass (Cu 90.9 + Zn 9.1)	3.64		0–100	.00204				
Brass (Cu 65.8 + Zn 34.2)	6.29		0–100	.00158				
Bronze (Cu 88 + Sn 12)	17.8		19–92	.0005				
Cadmium (drawn)		7.60	20	.0038				
Cesium	19.0							
Calcium (99.57% pure)		4.59		.0036				
Calido * (Ni 59 + Cr 16 + Fe 25)		110	20	.00025	1,000	8.15	90,000	16
Calorite * (Ni 65 + Cr 12 + Fe 15 + Mn 8)								
Carbon (graphite)	(800–1,300)		25–387	(−.0006 −.0012)	400	1.56	500	
Carbon (incandescent lamp)	4,000		25–335	−.0003	350	1.54	600	
Chromax * (Ni 30 + Cr 20 + Fe 50)		100	20–500	.00031	5,100	7.95	70,000	15.8
Chromel * A (Ni 80 + Cr 20)		108	20–500	.00013	1,100	8.4	95,000	17
Chromel * C (Ni 60 + Cr 16 + Fe 24)		112	20–500	.00017	900	8.24	95,000	17
Chromel * D (Ni 30 + Cr 20 + Fe 50)		99.5	20–500	.00032	500	7.94	70,000	16
Chromium	2.6							
Chronin * (Ni 83.7 + Cr 14.7)	(Similar Chromel A)							
Cimet * (Ni 25 + Fe 75)	(Similar -Phenix)							

Note a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

* Trademark names.

Table 2. Properties of Conducting Materials at Usual Temperatures—*Continued*

Description of Material	Resistivity Microhm-cm 0° C	Resistivity Microhm-cm 20° C	Temperature Coefficient, See Note a: Temp., °C	Temperature Coefficient, See Note a: α	Max. Working Temp., °C	Density, $\frac{gm}{cm^3}$	Tensile Str. (annealed), $\frac{lb}{in.^2}$	Coef. Lin. Expansion per °C $10^{-6}\times$
Climax * (Ni 25 + Fe 74 + Mn 1)		87	20	.0007				
Cobalt (99.8% pure)		9.7	0–100	.00658				
			0	.0033				
Comet * (Ni 30 + Cr 4.75 + Fe 65.25)		95	20–500	.00088	600	8.15	55,000	15
Constaloy * (Ni 45 + Cu 55)	(Similar	Advance)						
Constantan * (Cu 60 + Ni 40)	49.0	49	0–100	0.0000±			62,000	
			12	.000008				
			25	.000002				
			100	−.000033				
			200	−.000020				
			500	.000027				
Copel * (Ni 45 + Cu 55)		48.8	0–100	.00002	500	8.86	60,000	14.9
Copper (annealed standard)	1.589	1.7241	0–100	.00427	500		35,000	17
			20	.00393				
Copper (electrolytic)	1.56		0–100	.00428				
			100	.0038				
Copper (hard-drawn)	1.60	1.77	0–100	.00408	600	8.92	45,000	16.6
			20	.00382				
Copper-iron (Fe 0.4%)	4.08		0–100	.00155				
Copper-manganese (Cu 70 + Mn 30)	100		0–100	.00004				
Copper-manganese-iron (Cu 70.6 + Mn 23.2 + Fe 6.2)	77		0	.000022				
Copper-manganese-nickel (Cu 73 + Mn 24 + Ni 3)	48		0	−.00003				
Corronil * (Ni 70 + Cu 26 + Mn 4)	(Similar	Lucero)						
Cronin * D	(Similar	Comet)						
Cronit * (Ni 60 + Cr 40)								
Cupron *	(Similar	Advance)						
Dilver * (similar Invar)								
Dumet * (Ni 46 + Fe 54)	(Similar	Invar)						
Electris * (similar Phenix)		83.2	20	.0011		8.1		
Elinvar * (Ni 36 + Cr 12 + Fe 52)		8						
Eureka * (similar Lucero)	47		0	.00005				
Evanohm * (Cr 20 + Al 2.5 + Cu 2.5 + Ni bal.)		133	−50 to 100	±0.00002	900		100,000	14
Excello * (Ni 85 + Cr 14 + Fe 0.5 + Mn 0.5)	91.4	92	20	.00016	1,150			
Excelsior * (similar Advance)		49.2	20	.0000		8.9		
Ferro-nickel	27.1	28.2	20	.00207	340	7.8		
Gallium	53							
German silver † 18% (Ni 18 + Cu 64 + Zn 18)	33.1	33.8	20	.00031	260	8.5		17.3
Germanium	89,000							
Gold (99.9%)	2.22		18–100	.00368	500	19.3	20,000	14.2
			20	.0034				
Glowray * (Ni 65 + Cr 12 + Fe 23)		(Similar	Chromel	C)				

NOTE a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

* Trademark names.

† German Silver-30% has substantially the same properties as "Advance."

Table 2. Properties of Conducting Materials at Usual Temperatures—*Continued*

Description of Material	Resistivity Microhm-cm, 0° C	Resistivity Microhm-cm, 20° C	Temperature Coefficient, See Note a: Temp., °C	Temperature Coefficient, See Note a: α	Max. Working Temp., °C	Density, gm/cm³	Tensile Str. (annealed), lb/in.²	Coef. Lin. Expansion per °C $10^{-6}\times$
Hipernik * (Ni 50 + Fe 50)								
Hopkinson alloy * (Ni 25 + Fe 75)		(Similar	Phenix)					
Hytemco * (Ni 72 + Fe 28)		20	20–100	.0045	500	8.46	70,000	15
Ia Ia * (Ni 40 + Cu 60 soft)	47.1	47.1	20	.000005		8.92		
Ia Ia * (Ni 40 + Cu 60 hard-drawn)	50.2	50.2	20	.000011		8.92		
Ideal * (Ni 40 + Cu 58 + Fe 1 + Mn 1)	49		0–100	.0000±				
			20	.000005	520	8.9	65,000	14
Ilium		91.61		.000479				
Inconel * (Cr 13 + Fe 8 + Ni 49)	(Similar	Nirex)						
Indium	8.37		0	0.0047				
Invar * (Ni 36 + Fe 64)					150			1.5
Invariant * (Ni 47 + Fe 53)	(Similar-	Platinite)						As glass
Iridium	6.10		0–100	.00411				
Iron (pure)	8.85		0–100	.00625				
			100	.0068				
Iron (99.98% pure)		10	20	.0050	1,000	7.86	80,000	11.7
Iron (steel) soft	11.8		10–35	.00423				
Iron (steel) tempered glass hard	45.7		10–35	.00161				
			0	.0016				
Iron, cast (soft)	74.4							
Iron, cast (hard)	97.8							
Kanthal D (Cr 23 + Al 3 + Co2 + Fe bal.)		135	20–500	.00003	900		100,000	16
Karma * (Ni 80 + Cr 20)		103	20	.00016	1,100		90,000	15
Kromax * (similar Karma)								
Kromore * (Ni 85 + Cr 15)		94.6	20	.000242	1,100	8.9	80,000	
Krupp metal (nickel steel)	85.0			.00070				
Lead	19.8	22	0–100	.00411				
			20	.0039				
Lead-bismuth (Pb 42.3 + Bi 57.7)	63.3							
Lithium	8.55		0	.0047				
Lohm * (Ni 6 + Cu 94)		10	20–100	.00071	700	8.9	70,000	18
Lucero * (Ni 70 + Cu 30)		48.2	20–250	.0010	600	8.9	100,000	12.5
Magno * (Ni 95 + Mn 5)								
Magnesium		4.6	20	.004				
Magnesium (free from zinc)	4.35		0	.0038				
Manganese	5.0±							
Manganese-copper * (Cu 70 + Mn 30)	100		0–100	.00004				
Manganese-nickel (Mn 2 + Ni 98)		14	20–100	.0045	1,100	8.8	90,000	14.6
Manganin * (Cu 84 + Mn 12 + Ni 4)		48.2	15–35	.000015	100	8.2	60,000	18.7
			25	.000000				
			100	−.000042				
Magno * (Mn 4.5 + Ni 95.5)		20	20–100	.0036	1,100	8.75	90,000	14.3

Note a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

* Trademark names.

Table 2. Properties of Conducting Materials at Usual Temperatures—*Continued*

Description of Material	Resistivity Microhm-cm		Temperature Coefficient, See Note a		Max. Working Temp., °C	Density, $\frac{gm}{cm^3}$	Tensile Str. (annealed), $\frac{lb}{in.^2}$	Coef. Lin. Expansion per °C $10^{-6}\times$
	0° C	20° C	Temp., °C	α				
Marsh's patent * (Ni 75 + Cr 25)	(Similar Chr	omel A)						
Mercury	94.07	95.783	0–100	Note b				
			20	.00089				
Midohm * (Ni 23 + Cu 77)		30	20–100	.00018	700	8.9	70,000	17.5
Molybdenum (very pure)	5.14		0–100	.00435				
			25	.0033				
Molybdenum (annealed)	4.2		0–170	.0050				
Molybdenum (hard-drawn)	4.9		0–170	.0050				
Mond * No. 70 (Ni 70 + Cu 26 + Mn 4)	(Similar	Lucero)						
Monel metal (Ni 67 + Cu 28 + Mn 5)	40.8	42.6	20	.0020	425	8.15	44,000	12.6
Nichrome * (Ni 61 + Cr 15 + Fe 24)		112	20–500	.00017	980	8.25	120,000	17
Nichrome * V (Ni 80 + Cr 20)		108	20–500	0.00013	1,100	8.41	95,000	17
Nickel		7.8	20	.006				
Nickel (electrolytic)	6.93		0–100	.00618				
Nickel (very pure)		7.236						
Nickel (commercial wire)	9.9		20	.00400	700	8.8	70,000	14,0
Nickel-chromium (Ni + Cr and Fe + Mn)								
Nickel-silver 18% (Ni 18 + Cu 64 + Zn 18)		33.3	20	.00027	260	8.5	60,000	17.3
Nickel-silver 30% (Ni 30 + Cu 50 + Zn 20)		48.2	20	.00020	260	8.5	60,000	
Nickel steel (4.35% Ni)	29.4							
Nickelin * (same German silver)								
Nicraloy * (similar Chromel)								
Nilvar * (Ni 36 + Fe 64)		80.5	20–100	.00135	500	8.08	100,000	1.0
Nirex * (Cr 13 + Fe 8 + Ni 79)		98.1	20–500	.00012	1,100	8.55	130,000	16.1
Ohmax *		166	20–500	.000066	500	6.8	130,000	15.8
Osmium		9.5						
Palladium		11	20	.0033				
Palladium (very pure)	10.21..		0	.0035				
Peerless * (Ni 78.5 + Cr 16 + Fe 3 + Mn 2)		95.5	20	.00018	1,100	8,05		
Phenix * (Ni 25 + Fe 75)		83.1	20	.0011	400	8.10	75,000	14
Phosphor-bronze	7.75		0	.0040–				
Placet * (Ni 60 + Cr 15 + Fe 20 + Mn 5)	(Similar	Premier)						
Platinite * (Ni 42–46, Fe 58–54)	45		0	.003				
Platinoid * (Cu 62 + Ni 15 + Zn 22)		34.4 (18°)						
Platinum	9.83		20	.003				

NOTE a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

NOTE b: Use equation $R_t = R_0 (1 + aT + bT^2)$ with $a = 0.0008649$ and $b = 0.00000112$.

* Trademark names.

Table 2. Properties of Conducting Materials at Usual Temperatures—*Continued*

Description of Material	Resistivity Microhm-cm		Temperature Coefficient, See Note a		Max. Working Temp., °C	Density, gm/cm³	Tensile Str. (annealed), lb/in.²	Coef. Lin. Expansion per °C $10^{-6}\times$
	0° C	20° C	Temp., °C	α				
Platinum, drawn, wire	10.96		0–100	.00367	1,200	21.45	50,000	8.9
			0	.0037				
Platinum-iridium (Pt 80 + Ir 20)	31.6		−100 to +100	.002±				
			0	.0008				
Platinum-rhodium (Pt 90 + Rh 10)	21.14		15	.00143				
Potassium	6.1		0	.0057				
Premier * (Ni 61 + Cr 11 + Fe 25 + Mn 3)		103.0	20	.00036	1,000	8.15		
Radiohm * (Cr 16.5 + Al 5 + Fe 78.5)		133	20–500	.0001	500	7.30	90,000	15.5
Rayo * (Ni 85 + Cr 15)		95.7	20	.00018	1,100	8.05	90,000	15
Redray * (Ni 85 + Cr 15)	(Similar	Rayo)						
Rheotan * (Cu 84 + Fe 12 + Zn 4)	44.6		0	.00041				
Rheotan * II (Cu 53.3 + Ni 25.3 + Fe 4.5 + Zn 16.9)	53		0	0.0004				
Rhodium	5.11		0	.0043				
Rose's metal * (Bi 48.9 + Sn 23.5 + Pb 27.6)	64.5		0–94.3	.0023				
Rubidium	11.6		0	.0060				
R-63 Alloy (Mn 4 + Si 1 + Ni 95)		25	20–250	.0027	1,100	8.72	130,000	15.2
Silchrome * (Si + Cr + Fe)		113.0	20	.000025	1,100	7.63		
Silicon		58±						
Silver (99.78% pure)		1.629 (18°)	20	.0038	500	10.5	42,000	18.9
Silver (electrolytic)	1.468		0–100	.00400				
Sodium	4.3		0	.0054				
Stainless Type 304 (Cr 18 + Ni 8 + Fe 74)		73	20–500	.00094	1,100	7.93	200,000	20
Steel (see iron)								
Strontium		24.8						
Superior * (Ni 78 + Cr 19.5 + Fe 0.5 + Mn 2)		103.0	20	.00011	1,100	8.2		
Tantalum	14.6	15.5	0–100	.0033				
			20	.0031				
Tarnac * (similar Manganin)		41	20	.000025	100	8.89		
Tellurium		$.2\times10^6$ (19.6°)						
Thallium (pure)	17.6							
Therlo * (Cu + Mn + Al)	46.7	46.7	20	56×10^{-7}	200	8.15	78,000	19.4
Tico * (Ni 27.5 + Fe 72.5)	(Similar	Phenix)						
Tin	10.5	11.5	0–100	.00465				
			20	.0042				
Tophet * A	(Similar Chr	omel A)						
Tophet * C	(Similar Chr	omel C)						
Tophet * D	(Similar Chr	omel D)						
Tungsten		5.51	18	.0045	2,000	19.3	600,000	4

Note a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

* Trademark names.

Table 2. Properties of Conducting Materials at Usual Temperatures—*Continued*

Description of Material	Resistivity Microhm-cm		Temperature Coefficient, See Note a		Max. Working Temp., °C	Density, gm/cm³	Tensile Str. (annealed), lb/in.²	Coef. Lin. Expansion per °C $10^{-6}\times$
	0° C	20° C	Temp., °C	α				
Tungsten (annealed)		4.37	0–170	.0051				
Wood's metal * (Bi 55.7 + Sn 13.7 + Pb 13.7 + Cd 16.2)	51.8		0–69.8	.0023				
Yankee silver * (similar Nickel silver)		33.0	20	.000155		8.6		15.9
Zinc (pure)	5.38		18–100	.00402				
Zinc (trace Fe)	5.75	5.92	20	.00347	100	7.14	25,000	33
14 Alloy (Ni 42 + Cr 5.5 + Fe 52.5)		93.1	20–500	.0025	1,100	8.10	100,000	10.8
30 Alloy (Ni 2.25 + Cu 97.5)		5.0	20–100	.0013	500	8.9	30,000	17.5
42 Alloy (Ni 42 + Fe 58)		66.5	20–500	.0012	1,100	8.12	100,000	5.3
45 Alloy	(Similar	Advance)						
46 Alloy (Ni 46 + Fe 54)		45.7	20–500	.0027	1,100	8.17	100,000	8.0
52 Alloy (Ni 51 + Fe 49)		43.2	20–500	.0029	1,100	8.25	100,000	9.5
60 Alloy	(Similar	Lohm)						
90 Alloy (Ni 11 + Cu 89)		15	20–100	.00049	500	8.9	35,000	17.5
95 Alloy	(Similar	90 Alloy)						
99 Alloy (Ni 99.8)		8	0–100	.0060				
193 Alloy (Ni 30 + Cr 2 + Fe 67 + Mn 1)		87.2	20	.00072	650	8.15	60,000	17.1
331 Alloy	(Similar Eva	nohm)						13.1
525 Alloy		100	20–500	.00034	500			

Note a: Where a temperature *range* is given, the coefficient α is the mean value for the range, referred to the lower temperature.

* Trademark names.

Table 3. Properties of Conducting Materials at High Temperatures

Description of Material	Resistivity in Microhm centimeters		
	500° C	1000° C	1500° C
Aluminum (fused)		24	29
Aluminum (solid)	10		
Alundum		8×10^9	75×10^7
Antimony (b) fused		136	
Antimony (solid)	152		
Bismuth (fused)	139.9	167.5	
Boron	60×10^6 appx.		
Brass (2–1 fused)		41	
Brass (2–1 solid)	12.5		
Cadmium (fused)	34.12		
Calido (solid)	109	122	136
Carbon (a)	2700	2400	2200
Carbon (b)	3700	3400	2900
Carbon (c)	3300	3000	
Carbon (d)	2800	2100	1600
Carbon grains (a)	8.5×10^6 appx.	2.8×10^6	
Carbon grains (b)	4.8×10^6 appx.	1.9×10^6	0.85×10^6
Carbon powder	0.22×10^6	0.12×10^6	
Copper (fused)			24.8
Copper (solid)	5.1	9.42	
Copper chloride (fused)	2.50×10^6		
Copper oxide (Cu O)	5640×10^6		
Copper oxide (Cu O, powder)		18×10^6	
Copper oxide (Cu O_2 powder)	1570×10^6		
Ferro-nickel (solid)	94	105	
Glass	330×10^6 appx.	1×10^6 appx.	
Gold (fused)			37
Gold (solid)	6.62	12.54	
Graphite (a)	840	860	890
Graphite (b)	800	650	580
Graphite grains	2.70×10^6	1.7×10^6	1.2×10^6
Iron (a) solid	52 appx.	111 appx.	131 appx.
Iron (b) fused			166
Iron oxide (Fe_2O_3, powder)	1260×10^6	31.4×10^6	
Krupp metal (solid)	115		
Kryptol*	10×10^6	4.8×10^6	3.4×10^6
Lead (fused)	102.85	125	148
Lead chloride (fused 520°)	0.418×10^6		
Lead chloride (solid)	0.824×10^6		
Lead-tin alloy (fused)	81	98	
Magnesium oxide (powder)		1400×10^6	
Manganese oxide (powder)		15.7×10^6	
Manganese oxide (MnO_2, powder)	2200×10^6		
Molybdenum (solid)	16.5	28.5	40.5
Nernst filament			0.5×10^6 appx.
Nichrome* II (solid)	119	128	
Platinum (a) solid	34.4	66	98
Platinum (b) solid	25.3	40.8	52.6
Porcelain		15×10^6 appx.	
Quartz		110×10^6	
Refrax*	19.7×10^6	3.7×10^6	0.5×10^6

* Trademark names.

Table 3. Properties of Conducting Materials at High Temperatures—*Continued*

Description of Material	Resistivity in Microhm-centimeters		
	500° C	1000° C	1500° C
Silfrax* B	0.92×10^6	0.84×10^6	0.7×10^6
Silicon	0.094×10^6 to 0.023×10^6		
Silicon powder	120×10^6	3.5×10^6	
Silver (fused)		17.01	23
Silver (solid)	5		
Silver chloride (fused)	0.547×10^6		
Sodium chloride (fused)		0.90×10^6	
Tantalum (solid)	36	57	
Tantalum (a) solid			78
Tantalum (b) solid			74
Tin (fused)	54.62	68	80.5
Tungsten (solid)			43
Tungsten (a) solid	18	30.5	
Tungsten (b) solid	18	33.4	50
Zinc (fused)	36.60		
Zinc oxide (powder)		26.7×10^6	

* Trademark names.

Table 4. Approximate Values for the Physical Properties of Beryllium-copper Alloys of the 21 Per Cent Beryllium Class

Condition	Solution-treated "Annealed"	Solution-treated and Cold-worked	Solution-treated and Precipitation-hardened	Solution-treated, Cold-worked, and Precipitation-hardened
Electrical conductivity % I.A.-C.S. at 20° C	17	17	(*a*) 20–25 (*b*) 32–38	(*a*) 20–25 (*b*) 32–38
Tensile strength, psi	70,000	90,000	(*a*) 160,000 (*b*) 130,000	(*a*) 180,000 (*b*) 140,000
Yield strength, psi, at 0.5% elongation under load			(*a*) 150,000	(*a*) 175,000
Elongation in 2 in., %	35		(*a*) 3.0	(*a*) 2.0
Modulus of elasticity, psi	16×10^6		$18.4\text{–}19.4 \times 10^6$	
Endurance limit, psi, at 10^8 reversals of stress		23,000	28,000	28,000

(*a*) Heat-treated for maximum hardness. (*b*) Heat-treated for maximum conductivity.

3. WIRE TABLES

Table 5. Solid Copper Wire

A. W. G. or B. & S. Gage; English Units

100 per cent conductivity; density 8.89 at 20 deg cent

Gage No.	Diameter in Mils	Cross-section		Resistance at 20° C or 68° F		Weight in Pounds		Feet per Pound
		Circular Mils	Square Inches	Ohms per 1000 ft	Ohms per Mile	per 1000 ft	per Mile	
0000	460.0	211,600	0.1662	0.04901	0.259	640.5	3380	1.561
000	409.6	167,800	0.1318	0.06180	0.326	507.9	2680	1.968
00	364.8	133,100	0.1045	0.07793	0.411	402.8	2130	2.482
0	324.9	105,500	0.08289	0.90827	0.519	319.5	1680	3.130
1	289.3	83,690	0.06573	0.1239	0.654	253.3	1340	3.947
2	257.6	66,370	0.05213	0.1563	0.825	200.9	1060	4.977
3	229.4	52,640	0.04134	0.1970	1.04	159.3	841	6.276
4	204.3	41,740	0.03278	0.2485	1.31	126.4	667	7.914
5	181.9	33,100	0.02600	0.3133	1.65	100.2	529	9.980
6	162.0	26,250	0.02062	0.3951	2.09	79.46	420	12.58
7	144.3	20,820	0.01635	0.4982	2.63	63.02	333	15.87
8	128.5	16,510	0.01297	0.6282	3.32	49.98	264	20.01
10	101.9	10,380	0.008155	0.9989	5.28	31.43	166	31.82
12	80.81	6,530	0.005129	1.588	8.38	19.77	104	50.59
14	64.08	4,107	0.003225	2.525	13.3	12.43	63.3	80.44
15	57.07	3,257	0.002558	3.184	16.8	9.858	52.0	101.4
16	50.82	2,583	0.002028	4.015	21.2	7.818	41.3	127.9
17	45.26	2,048	0.001609	5.064	26.7	6.200	32.7	161.3
18	40.30	1,624	0.001276	6.385	33.7	4.917	26.0	203.4
19	35.89	1,288	0.001012	8.051	42.5	3.899	20.6	256.5
20	31.96	1,022	0.0008023	10.15	53.6	3.092	16.3	323.4
21	28.46	810.1	0.0006363	12.80	67.6	2.452	12.9	407.8
22	25.35	642.4	0.0005046	16.14	85.2	1.945	10.3	514.2
23	22.57	509.5	0.0004002	20.36	108	1.542	8.14	648.4
24	20.10	404.0	0.0003173	25.67	135	1.223	6.46	817.7
25	17.90	320.4	0.0002517	32.37	171	0.9699	5.12	1,031
26	15.94	254.1	0.0001996	40.82	216	0.7692	4.06	1,300
27	14.20	201.5	0.0001583	51.46	272	0.6100	3.22	1,639
28	12.64	159.8	0.0001255	64.90	343	0.4837	2.55	2,067
29	11.26	126.7	0.00009953	81.84	432	0.3836	2.03	2,607
30	10.03	100.5	0.00007894	103.2	545	0.3042	1.61	3,287
31	8.928	79.70	0.00006260	130.1	687	0.2413	1.27	4,145
32	7.950	63.21	0.00004964	164.1	866	0.1913	1.01	5,227
33	7.080	50.13	0.00003937	206.9	1,090	0.1517	0.814	6,591
34	6.305	39.75	0.00003122	260.9	1,380	0.1203	0.635	8,310
35	5.615	31.52	0.00002476	329.0	1,740	0.09542	0.504	10,480
36	5.000	25.00	0.00001964	414.8	2,190	0.07568	0.400	13,210
37	4.453	19.83	0.00001557	523.1	2,762	0.06001	0.317	16,660
38	3.965	15.72	0.00001235	659.6	3,480	0.04759	0.251	21,010
39	3.531	12.47	0.000009793	831.8	4,392	0.03774	0.199	26,500
40	3.145	9.888	0.000007766	1049	5,540	0.02993	0.158	33,410
41	2.800	7.842	0.000006159	1323	6,983	0.02374	0.125	42,130
42	2.494	6.219	0.000004884	1668	8,806	0.01882	0.0994	53,120
43	2.221	4.932	0.000003873	2103	11,100	0.01493	0.0788	66,990
44	1.978	3.911	0.000003072	2652	14,000	0.01184	0.0625	84,470

Table 6. Solid Copper Wire
A. W. G. or B. & S. Gage in Metric Units
100 per cent conductivity; density 8.89 at 20 deg cent

Gage No.	Diameter, mm	Cross-section, sq mm	Ohms per Kilometer 20° C	Kilograms per Kilometer
0000	11.68	107.2	0.1608	953.2
000	10.40	85.03	0.2028	755.9
00	9.266	67.43	0.2557	599.5
0	8.252	53.48	0.3224	475.4
1	7.348	42.41	0.4066	377.0
2	6.544	33.63	0.5126	299.0
3	5.827	26.67	0.6464	237.1
4	5.189	21.15	0.8152	188.0
5	4.621	16.77	1.028	149.1
6	4.115	13.30	1.296	118.2
7	3.665	10.55	1.634	93.78
8	3.264	8.366	2.061	74.37
10	2.588	5.261	3.277	46.77
12	2.053	3.309	5.211	29.42
14	1.628	2.081	8.285	18.50
15	1.450	1.650	10.45	14.67
16	1.291	1.309	13.18	11.63
17	1.150	1.038	16.61	9.226
18	1.024	0.8231	20.95	7.317
19	0.9116	0.6527	26.42	5.803
20	0.8118	0.5176	33.31	4.602
21	0.7230	0.4105	42.00	3.649
22	0.6438	0.3255	52.96	2.894
23	0.5733	0.2582	66.79	2.295
24	0.5106	0.2047	84.22	1.820
25	0.4547	0.1624	106.2	1.443
26	0.4049	0.1288	133.9	1.145
27	0.3606	0.1021	168.8	0.9078
28	0.3211	0.08098	212.9	0.7199
29	0.2859	0.06422	268.5	0.5709
30	0.2546	0.05093	338.6	0.4527
31	0.2268	0.04039	426.9	0.3590
32	0.2019	0.03203	538.3	0.2847
33	0.1798	0.02540	678.8	0.2258
34	0.1601	0.02014	856.0	0.1791
35	0.1426	0.01597	1079	0.1420
36	0.1270	0.01267	1361	0.1126
37	0.1131	0.01005	1716	0.08931
38	0.1007	0.007967	2164	0.07083
39	0.08969	0.006318	2729	0.05617
40	0.07987	0.005010	3441	0.04454
41	0.07113	0.003973	4339	0.03532
42	0.06334	0.003151	5472	0.02801
43	0.05641	0.002499	6900	0.02222
44	0.05023	0.001982	8700	0.01762

Table 7. Solid Copper Wire
British Standard Wire Gage; English Units
100 per cent conductivity; density 8.89 at 20 deg cent

Gage No.	Diameter, mils	Cross-section		Ohms per 1000 ft, 15.6° C or 60° F *	Pounds per 1000 ft
		Circular Mils	Square Inches		
7–0	500	250,000	0.1964	0.04077	756.8
6–0	464	215,300	0.1691	0.04734	651.7
5–0	432	186,600	0.1466	0.05461	564.9
4–0	400	160,000	0.1257	0.06370	484.3
3–0	372	138,400	0.1087	0.07365	418.9
2–0	348	121,100	0.09512	0.08416	366.6
0	324	105,000	0.08245	0.09709	317.8
1	300	90,000	0.07069	0.1132	272.4
2	276	76,180	0.05983	0.1338	230.6
3	252	63,500	0.04988	0.1605	192.2
4	232	53,820	0.04227	0.1894	162.9
5	212	44,940	0.03530	0.2268	136.0
6	192	36,860	0.02895	0.2765	111.6
7	176	30,980	0.02433	0.3290	93.76
8	160	25,600	0.02011	0.3981	77.49
9	144	20,740	0.01629	0.4915	62.77
10	128	16,380	0.01287	0.6221	49.59
11	116	13,460	0.01057	0.7574	40.73
12	104	10,820	0.008495	0.9423	32.74
13	92	8,464	0.006648	1.204	25.62
14	80	6,400	0.005027	1.592	19.37
15	72	5,184	0.004072	1.966	15.69
16	64	4,096	0.003217	2.488	12.40
17	56	3,136	0.002463	3.250	9.493
18	48	2,304	0.001810	4.424	6.974
19	40	1,600	0.001257	6.370	4.843
20	36	1,296	0.001018	7.864	3.923
22	28	784.0	0.0006158	13.00	2.373
24	22	484.0	0.0003801	21.06	1.465
26	18	324.0	0.0002545	31.46	0.9807
28	14.8	219.0	0.0001720	46.54	0.6630
30	12.4	153.8	0.0001208	66.28	0.4654
32	10.8	116.6	0.00009161	87.38	0.3531
34	9.2	84.64	0.00006648	120.4	0.2562
36	7.6	57.76	0.00004536	176.5	0.1748
38	6.0	36.00	0.00002827	283.1	0.1090
40	4.8	23.04	0.00001810	442.4	0.06974
42	4.0	16.00	0.00001257	637.0	0.04843
44	3.2	10.24	0.000008042	995.3	0.03100
46	2.4	5.760	0.000004524	1,769	0.01744
48	1.6	2.560	0.000002011	3,981	0.007749
50	1.0	1.000	0.0000007854	10,190	0.003027

* Let C = per cent conductivity, R_{60} = resistance of 100 per cent conductivity wire at 60 deg fahr (from table), R_t = resistance of wire of conductivity C at any temperature t deg fahr; then

$$R_t = \frac{100}{C} R_{60} [1 + 0.00223(t - 60)]$$

Table 8. Solid Copper Wire

"Millimeter Gage"; Metric Units and Circular Mils

100 per cent conductivity; density 8.89 at 20 deg cent

Diameter, mm	Cross-section, sq mm	Ohms per Kilometer, 20° C	Kilograms per Kilometer	Cross-section, cir mils *
10.0	78.54	0.2195	698.2	155,000
9.0	63.62	0.2710	565.6	125,550
8.0	50.27	0.3430	446.9	99,200
7.0	38.48	0.4480	342.1	75,950
6.0	28.27	0.6098	251.4	55,800
5.0	19.64	0.8781	174.6	38,750
4.5	15.90	1.084	141.4	31,380
4.0	12.57	1.372	111.7	24,800
3.5	9.621	1.792	85.53	18,990
3.0	7.069	2.439	62.84	13,950
2.5	4.909	3.512	43.64	9,690
2.0	3.142	5.488	27.93	6,200
1.8	2.545	6.775	22.62	5,010
1.6	2.011	8.575	17.87	3,970
1.4	1.539	11.20	13.69	3,040
1.2	1.131	15.24	10.05	2,230
1.0	0.7854	21.95	6.982	1,550
0.90	0.6362	27.10	5.656	....
0.80	0.5027	34.30	4.469	....
0.70	0.3848	44.80	3.421	....
0.60	0.2827	60.98	2.514	....
0.50	0.1964	87.81	1.746	....
0.45	0.1590	108.4	1.414	....
0.40	0.1257	137.2	1.117	....
0.35	0.09621	179.2	0.8553	....
0.30	0.07069	243.9	0.6284	....
0.25	0.04909	351.2	0.4364	....
0.20	0.03142	548.8	0.2793	....
0.15	0.01767	975.6	0.1571	....
0.10	0.007854	2195	0.06982	....
0.05	0.001964	8781	0.01746	....

* One square millimeter equals 1973.52 circular mils.

Table 9. Solid Copper Wire; Ohms per Unit Weight

A. W. G. or B. & S. Gage; English and Metric Units

100 per cent conductivity; density 8.89 at 20 deg cent

Gage No.	Ohms per Pound			Ohms per Kilogram		
	0° C 32° F	20° C 68° F	50° C 122° F	0° C	20° C	50° C
0000	0.00007051	0.00007652	0.00008554	0.0001554	0.0001687	0.0001886
000	0.0001121	0.0001217	0.0001360	0.0002472	0.0002682	0.0002999
00	0.0001783	0.0001935	0.0002163	0.0003930	0.0004265	0.0004768
0	0.0002835	0.0003076	0.0003439	0.0006249	0.0006782	0.0007582
1	0.0004507	0.0004891	0.0005468	0.0009936	0.001078	0.001206
2	0.0007166	0.0007778	0.0008695	0.001580	0.001715	0.001917
3	0.001140	0.001237	0.001383	0.002512	0.002726	0.003048
4	0.001812	0.001966	0.002198	0.003995	0.004335	0.004846
5	0.002881	0.003127	0.003495	0.006352	0.006893	0.007706
6	0.004581	0.004972	0.005558	0.01010	0.01096	0.01225
7	0.007284	0.007906	0.008838	0.01606	0.01743	0.01948
8	0.01158	0.01257	0.01405	0.02553	0.02771	0.03098
9	0.01842	0.01999	0.02234	0.04060	0.04407	0.04926
10	0.02928	0.03178	0.03553	0.06456	0.07006	0.07833
11	0.04656	0.05053	0.05649	0.1026	0.1114	0.1245
12	0.07404	0.08035	0.08983	0.1632	0.1771	0.1980
13	0.1177	0.1278	0.1428	0.2595	0.2817	0.3149
14	0.1872	0.2032	0.2271	0.4127	0.4479	0.5007
15	0.2976	0.3230	0.3611	0.6562	0.7121	0.7961
16	0.4733	0.5136	0.5742	1.043	1.132	1.266
17	0.7525	0.8167	0.9130	1.659	1.800	2.013
18	1.197	1.299	1.452	2.638	2.863	3.201
19	1.903	2.065	2.308	4.194	4.552	5.089
20	3.025	3.283	3.670	6.670	7.238	8.092
21	4.810	5.221	5.836	10.60	11.51	12.87
22	7.649	8.302	9.280	16.86	18.30	20.46
23	12.16	13.20	14.76	26.81	29.10	32.53
24	19.34	20.99	23.46	42.63	46.27	51.73
25	30.75	33.37	37.31	67.79	73.57	82.25
26	48.89	53.06	59.32	107.8	117.0	131.8
27	77.74	84.37	94.32	171.4	186.0	207.9
28	123.6	134.2	150.0	272.5	295.8	330.6
29	196.6	213.3	238.5	433.3	470.3	525.7
30	312.5	339.2	379.2	689.0	747.8	836.0
31	497.0	539.3	602.9	1,096	1,189	1,329
32	790.2	857.6	958.7	1,742	1,891	2,114
33	1,256	1,364	1,524	2,770	3,006	3,361
34	1,998	2,168	2,424	4,404	4,780	5,344
35	3,177	3,448	3,854	7,003	7,601	8,497
36	5,051	5,482	6,128	11,140	12,080	13,510
37	8,032	8,717	9,744	17,710	19,220	21,480
38	12,770	13,860	15,490	28,150	30,560	34,160
39	20,310	22,040	24,640	44,770	48,590	54,310
40	32,290	35,040	39,170	71,180	77,260	86,360
41	51,340	55,720	62,290	113,200	122,800	137,300
42	81,640	88,600	99,050	180,000	195,300	218,400
43	129,800	140,900	157,500	286,200	310,600	347,200
44	206,400	224,000	250,400	455,000	493,900	552,100

Table 10. Solid Aluminum Wire

A. W. G. or B. & S. Gage; English Units

61 per cent conductivity; density 2.70

Gage No.	Diameter, mils	Cross-section: Circular Mils	Cross-section: Square Inches	Resistance at 20° C or 68° F *: Ohms per 1000 ft	Resistance at 20° C or 68° F *: Ohms per Mile	Weight in Pounds: per 1000 ft	Weight in Pounds: per Mile	Feet per Pound
0000	460.0	211,600	0.1662	0.0804	0.424	195	1027	5.14
000	409.6	167,800	0.1318	0.101	0.535	154	815	6.48
00	364.8	133,100	0.1045	0.128	0.675	122	646	8.17
0	324.9	105,500	0.08289	0.161	0.851	97.0	512	10.31
1	289.3	83,690	0.06573	0.203	1.073	76.9	406	13.00
2	257.6	66,370	0.05213	0.256	1.353	61.0	322	16.39
3	229.4	52,630	0.04134	0.323	1.706	48.4	255	20.7
4	204.3	41,740	0.03278	0.408	2.15	38.4	203	26.1
5	181.9	33,100	0.02600	0.514	2.71	30.4	160.7	32.9
6	162.0	26,250	0.02062	0.648	3.42	24.1	127.4	41.4
7	144.3	20,820	0.01635	0.817	4.31	19.1	101.0	52.3
8	128.5	16,510	0.01297	1.03	5.44	15.2	80.2	65.9
10	101.9	10,380	0.008155	1.64	8.65	9.55	50.4	104.8
12	80.81	6,530	0.005129	2.61	13.76	6.00	31.7	166.6
14	64.08	4,107	0.003225	4.14	21.9	3.78	19.93	265
15	57.07	3,257	0.002558	5.22	27.6	2.99	15.81	334
16	50.82	2,583	0.002029	6.59	34.8	2.37	12.54	421
17	45.26	2,048	0.001609	8.31	43.8	1.88	9.94	531
18	40.30	1,624	0.001276	10.5	55.3	1.49	7.89	670
19	35.89	1,288	0.001012	13.2	69.7	1.18	6.25	844
20	31.96	1,022	0.0008023	16.7	87.9	0.939	4.96	1,065
21	28.46	810.1	0.0006363	21.0	110.9	0.745	3.93	1,343
22	25.35	642.4	0.0005046	26.5	139.8	0.591	3.12	1,693
23	22.57	509.5	0.0004002	33.4	176.3	0.468	2.47	2,130
24	20.10	404.0	0.0003173	42.1	222	0.371	1.961	2,690
25	17.90	320.4	0.0002517	53.1	280	0.295	1.556	3,390
26	15.94	254.1	0.0001996	67.0	353	0.234	1.233	4,280
27	14.20	201.5	0.0001583	84.4	446	0.185	0.978	5,400
28	12.64	159.8	0.0001255	106	562	0.147	0.776	6,810
29	11.26	126.7	0.00009953	134	709	0.117	0.615	8,580
30	10.03	100.5	0.00007894	169	894	0.0924	0.488	10,820
31	8.928	79.70	0.00006260	213	1127	0.0733	0.387	13,650
32	7.950	63.21	0.00004964	269	1421	0.0581	0.307	17,210
33	7.080	50.13	0.00003937	339	1792	0.0461	0.243	21,700
34	6.305	39.75	0.00003122	428	2260	0.0365	0.1929	27,400
35	5.615	31.52	0.00002476	540	2850	0.0290	0.1530	34,510

* Let C = per cent conductivity, R_{20} = resistance of 61 per cent conductivity wire at 20 deg cent (from table), R_t = resistance of wire of conductivity C at any temperature t deg cent; then

$$R_t = \frac{61\, R_{20}}{C}[1 + 0.004(t - 20)]$$

Table 11. Solid Aluminum Wire

A. W. G. or B. & S. Gage in Metric Units

61 per cent conductivity; density 2.70; temperature 20 deg cent or 68 deg fahr *

Gage No.	Diameter, mm	Cross-section, sq mm	Ohms per Kilometer	Kilograms per Kilometer
0000	11.68	107.2	0.264	289
000	10.40	85.03	0.333	230
00	9.266	67.43	0.419	182
0	8.252	53.48	0.529	144
1	7.348	42.41	0.667	114
2	6.544	33.63	0.841	90.8
3	5.827	26.67	1.06	72.0
4	5.189	21.15	1.34	57.1
5	4.621	16.77	1.69	45.3
6	4.115	13.30	2.13	35.9
7	3.665	10.55	2.68	28.5
8	3.264	8.366	3.38	22.6
10	2.588	5.261	5.38	14.2
12	2.053	3.309	8.55	8.93
14	1.628	2.081	13.6	5.62
15	1.450	1.650	17.1	4.46
16	1.291	1.309	21.6	3.53
17	1.150	1.038	27.3	2.80
18	1.024	0.8231	34.4	2.22
19	0.9116	0.6527	43.3	1.76
20	0.8118	0.5176	54.6	1.40
21	0.7230	0.4105	68.9	1.11
22	0.6438	0.3255	86.9	0.879
23	0.5733	0.2582	110	0.697
24	0.5106	0.2047	138	0.553
25	0.4547	0.1624	174	0.438
26	0.4049	0.1288	220	0.348
27	0.3606	0.1021	277	0.276
28	0.3211	0.08098	349	0.219
29	0.2859	0.06422	440	0.173
30	0.2546	0.05093	555	0.138
31	0.2268	0.04039	700	0.109
32	0.2019	0.03203	883	0.0865
33	0.1798	0.02540	1110	0.0686
34	0.1601	0.02014	1400	0.0544
35	0.1426	0.01597	1770	0.0431

* Let C = per cent conductivity, R_{20} = resistance of 61 per cent conductivity wire at 20 deg cent (from table), R_t = resistance of wire of conductivity C at any temperature t deg cent; then

$$R_t = \frac{61\ R_{20}}{C}[1 + 0.004(t - 20)]$$

The temperature coefficient is approximate only.

Table 12. Solid Steel Wire
American Steel Wire Gage; English Units
12.5 per cent conductivity; density 7.78

Am. Steel Wire Gage No.	Diameter		Cross-section		Resistance at 20° C or 68° F *		Weight in Pounds		Feet per Pound
	In.	Mils	Circular Mils	Square Inches	Ohms per 1000 ft	Ohms per Mile	per 1000 ft	per Mile	
	1/2	500.0	250,000	0.1964	0.332	1.752	662.5	3499	1.51
7-0		490.0	240,100	0.1886	0.346	1.825	636.3	3360	1.57
	15/32	468.8	219,800	0.1726	0.378	1.993	582.4	3075	1.72
6-0		460.0	211,600	0.1662	0.392	2.07	560.8	2961	1.78
	7/16	437.5	191,400	0.1503	0.433	2.29	507.2	2678	1.97
5-0		430.0	184,900	0.1452	0.449	2.37	490.0	2587	2.04
	13/32	406.3	165,000	0.1296	0.503	2.65	436.8	2306	2.28
4-0		393.8	155,100	0.1218	0.535	2.82	411.9	2175	2.42
	3/8	375.0	140,600	0.1104	0.590	3.12	372.6	1967	2.68
3-0		362.5	131,400	0.1032	0.631	3.33	348.2	1839	2.87
	11/32	343.8	118,200	0.09280	0.702	3.71	313.1	1653	3.19
2-0		331.0	109,600	0.08605	0.757	4.00	290.3	1533	3.44
	5/16	312.5	97,660	0.07670	0.850	4.49	258.8	1366	3.86
0		306.5	93,940	0.07378	0.883	4.66	249.0	1315	4.02
1		283.0	80,090	0.06290	1.036	5.47	212.2	1121	4.71
	9/32	281.3	79,100	0.06213	1.049	5.54	209.6	1107	4.77
2		262.5	68,910	0.05412	1.204	6.36	182.6	964.1	5.48
	1/4	250.0	62,500	0.04909	1.328	7.01	165.6	874.5	6.04
3		243.7	59,490	0.04665	1.397	7.38	157.4	831.0	6.35
4		225.3	50,760	0.03987	1.635	8.63	134.5	710.2	7.43
	7/32	218.8	47,850	0.03758	1.734	9.15	126.8	669.5	7.89
5		207.0	42,850	0.03365	1.936	10.22	113.6	599.5	8.81
6		192.0	36,860	0.02895	2.25	11.88	97.7	515.8	10.23
	3/16	187.5	35,160	0.02761	2.36	12.46	93.2	491.9	10.73
7		177.0	31,330	0.02461	2.65	13.98	83.0	438.4	12.04
8		162.0	26,240	0.02061	3.16	16.69	69.6	367.2	14.38
	5/32	156.3	24,410	0.01917	3.40	17.95	64.7	341.6	15.46
9		148.3	21,990	0.01727	3.77	19.92	58.3	307.8	17.16
10		135.0	18,200	0.01431	4.55	24.0	48.3	255.0	20.70
	1/8	125.0	15,630	0.01227	5.31	28.0	41.4	218.6	24.15
11		120.5	14,520	0.01140	5.71	30.2	38.5	203.2	25.98
12		105.5	11,130	0.00874	7.45	39.4	29.5	155.7	33.90
	3/32	93.8	8,789	0.00690	9.44	49.8	23.3	123.0	42.94
13		91.5	8,372	0.00658	9.91	52.3	22.1	117.2	45.16
14		80.0	6,400	0.00503	12.96	68.5	17.0	89.55	58.97
15		72.0	5,184	0.00407	16.01	84.5	13.7	72.53	72.80
16		62.5	3,906	0.00307	21.2	112.1	10.4	54.66	96.60
	1/16	62.5	3,906	0.00307	21.2	112.1	10.4	54.66	96.60
17		54.0	2,916	0.00229	28.5	150.2	7.73	40.80	129.5
18		47.5	2,256	0.00177	36.8	194.2	5.98	31.57	167.2
19		41.0	1,681	0.00132	49.4	261	4.45	23.52	224.4
20		34.8	1,211	0.00095	68.5	362	3.21	16.95	311.5
21		31.8	1,008	0.00079	82.3	435	2.67	14.11	374.4
	1/32	31.3	977	0.00076	85.0	449	2.59	13.66	386.5
22		28.6	818	0.00064	101.4	536	2.17	11.45	461.1
23		25.8	666	0.00052	124.6	658	1.76	9.31	567.0
24		23.0	529	0.00042	156.8	828	1.40	7.40	713.5
25		20.4	416	0.00033	199.4	1053	1.10	5.82	907.0

* Let C = per cent conductivity,
R_{20} = resistance of 12.5 per cent conductivity wire at 20 deg cent (from table),
R_t = resistance of wire of conductivity C at any temperature t deg cent; then

$$R_t = \frac{12.5\, R_{20}}{C}\,[1 + 0.006(t - 20)]$$

The temperature coefficient is approximate only.

Table 12. Solid Steel Wire—*Continued*

American Steel Wire Gage; English Units

12.5 per cent conductivity; density 7.78

Am. Steel Wire Gage No.	Diameter		Cross-section		Resistance at 20° C or 68° F *		Weight in Pounds		Feet per Pound
	In.	Mils	Circular Mils	Square Inches	Ohms per 1000 ft	Ohms per Mile	per 1000 ft	per Mile	
26		18.1	328	0.00026	253	1337	0.87	4.58	1152
27		17.3	299	0.00024	277	1464	0.79	4.19	1261
28		16.2	262	0.00021	316	1669	0.70	3.67	1438
29		15.0	225	0.00018	469	1947	0.60	3.15	1677
30		14.0	196	0.00015	424	2240	0.52	2.74	1925
31		13.2	174	0.00014	476	2510	0.46	2.44	2166
32		12.8	164	0.00013	506	2670	0.43	2.30	2303
33		11.8	139	0.00011	596	3150	0.37	1.95	2710
34		10.4	108	0.00008	767	4050	0.29	1.51	3489
35		9.5	90	0.00007	919	4850	0.24	1.26	4193
36		9.0	81	0 00006	1023	5410	0.21	1.13	4659

* Let C = per cent conductivity,
R_{20} = resistance of 12.5 per cent conductivity wire at 20 deg cent (from table),
R_t = resistance of wire of conductivity C at any temperature t deg cent; then

$$R_t = \frac{12.5\,R_{20}}{C}[1 + 0.006(t - 20)]$$

The temperature coefficient is approximate only.

COPPER-CLAD STEEL WIRE. This wire consists of a steel core and a concentric coat of copper permanently welded thereto. It is used chiefly for long-span transmission and telephone wire. It is made in several grades, which differ in the relative amounts of steel and copper. The grades are designated by the corresponding conductivity expressed as percentages of the Annealed Copper Standard: e. g., 40 per cent grade has a conductivity of 40 per cent.

Table 13. Copper-clad Steel Wire

A. W. G. or B. & S. Gage; English Units

40 per cent conductivity; density 8.26

Gage No.	Diameter, mils	Cross-section		Resistance at 23.9° C or 75° F *		Weight in Pounds		Feet per Pound
		Circular Mils	Square Inches	Ohms per 1000 ft	Ohms per Mile	per 1000 ft	per Mile	
0000	460.0	211,600	0.1662	0.123	0.649	595	3140	1.68
000	409.6	167,800	0.1318	0.154	0.813	471	2490	2.12
00	364.8	133,100	0.1045	0.195	1.03	374	1970	2.67
0	324.9	105,500	0.08289	0.246	1.30	297	1570	3.37
1	289.3	83,690	0.06573	0.310	1.64	235	1240	4.26
2	257.6	66,370	0.05213	0.390	2.06	186	982	5.38
3	229.4	52,630	0.04134	0.492	2.60	148	781	6.76
4	204.3	41,740	0.03278	0.622	3.28	117	618	8.55
5	181.9	33,100	0.02600	0.782	4.13	92.9	491	10.76
6	162.0	26,250	0.02062	0.987	5.21	73.7	389	13.57
7	144.3	20,820	0.01635	1.25	6.60	58.5	309	17.09
8	128.5	16,510	0.01297	1.57	8.29	46.4	245	21.6
9	114.4	13,090	0.01028	1.98	10.5	36.8	194	27.2
10	101.9	10,380	0.008155	2.50	13.2	29.2	154	34.2
11	90.74	8,234	0.006467	3.15	16.6	23.1	122	43.3
12	80.81	6,530	0.005129	3.97	21.0	18.3	96.6	54.6
13	71.96	5,178	0.004067	5.00	26.4	14.6	77.1	68.5
14	64.08	4,107	0.003225	6.31	33.3	11.5	60.7	87.0

* Let C = per cent conductivity,
$R_{23.9}$ = resistance of 40 per cent conductivity wire at 23.9 deg cent (from table),
R_t = resistance of wire of conductivity C at temperature t deg cent; then

$$R_t = \frac{40\,R_{23.9}}{C}[1 + 0.00432(t - 23.9)]$$

The temperature coefficient is approximate only.

ALLOY WIRES OF HIGH TENSILE STRENGTH. Copper alloys having a low conductivity but a tensile strength from 50 to 100 per cent greater than that of copper are sometimes used where strength or hardness is a primary requisite, as in long spans of small wires or for trolley wires.

TENSILE BREAKING LOAD. The tensile strength in pounds for solid wires from $^1/_{16}$ to $^1/_2$ in. in diameter is given in Table 14.

Table 14. Breaking Load for Solid Wires in Pounds per Wire

Gage No. A.W.G. or B. & S.	Diameter		Hard-drawn Copper (A.S.T.M.) *	Hard-drawn Aluminum (23,000 to 33,300 lb per sq in.)	Copper-clad Steel, 40 per cent Grade	Steel (100,000 lb per sq in.) †
	In.	Mils				
	$^1/_2$	500	9310	4520	11,400	19,640
0000		460	8140	3820	10,000	16,620
	$^7/_{16}$	437	7500	3460	9,250	15,030
000		410	6720	3030	8,300	13,180
	$^3/_8$	375	5800	2540	7,150	11,040
00		365	5540	2400	6,850	10,450
0		325	4520	1910	5,700	8,289
	$^5/_{16}$	312	4220	1770	5,400	7,670
1		289	3680	1530	4,800	6,573
2		258	3000	1240	4,000	5,213
	$^1/_4$	250	2830	1170	3,780	4,909
3		229	2420	1000	3,200	4,134
4		204	1950	810	2,600	3,278
	$^3/_{16}$	187	1680	693	2,300	2,761
5		182	1570	655	2,200	2,600
6		162	1270	532	1,800	2,062
7		144	1020	432	1,450	1,635
8		129	822	351	1,200	1,297
	$^1/_8$	125	780	335	1,150	1,227
9		114	660	287	975	1,028
10		102	528	234	800	816
11		91	423	191	650	647
12		81	337	155	510	513
13		72	268	126	410	407
14		64	213	103	330	323
	$^1/_{16}$	62	203	98	310	307

* Tensile strength in pounds per square inch ranging from 49,000 for No. 0000 to 66,200 for No. 14; see below.

† For wires having a tensile strength of S pounds per square inch, multiply by $S/100{,}000$. The tensile strength of steel varies from 60,000 to 225,000 lb per sq in.

INSULATING MATERIALS

By W. R. Dohan

4. DIELECTRIC PROPERTIES

Dielectric Constant. The dielectric constant, K, of an insulating material is the ratio of the capacitance of an electrode system using the material as a dielectric to its capacitance with a vacuum dielectric. When the material has appreciable losses so that the parallel capacitance, C_p (determined by comparison with a standard capacitor and a parallel conductance, G), differs significantly from the effective series capacitance C_s (determined by comparison with a standard capacitor and a series resistance, R), it is standard to use the value of C_p in computing the dielectric constant. These capacitances are related by the following formula:

$$C_p = \frac{C_s}{1 + \tan^2 \delta} = \frac{C_s}{1 + D^2} = C_s \cos^2 \delta$$

where δ = the loss angle or phase difference.
D = the dissipation factor.

Since the dielectric constant may vary considerably with frequency or temperature the conditions of measurement should always be stated.

Phase Difference or Loss Angle. The difference between the theoretical 90 electrical degrees phase advance of the current through a perfect capacitor and the actual angle of phase advance, θ, of the current through the dielectric material is known as the phase difference, or loss angle, δ.

Dissipation Factor. For convenience in reference and calculation the tangent of the loss angle, δ, has been assigned the standard designation "dissipation factor" and given the symbol D.

Power Factor. The power factor of a dielectric material is the ratio of the power loss in the material to the product of the applied voltage and current. The power factor therefore is equal to the cosine of the angle of phase advance, θ, and to the sine of the complementary angle, δ. The sine and tangent of a small loss angle are very nearly the same, so that the power factor is substantially equal to the dissipation factor for values less than 0.1.

Loss Factor. The product of the dielectric constant and the dissipation factor (or power factor if the value is less than 0.1) is known as the loss factor.

Power Loss. The power loss in a dielectric may be considered to take place in a fictitious shunt or series resistance, depending on the material, the applied frequency, and the temperature. For a series resistance, R, the dissipation factor is

$$D = \tan \delta = R2\pi fC_s \doteq \text{Power factor}$$

For a parallel conductance, G, or shunt resistance, r, the dissipation factor is

$$D = \tan \delta = \frac{G}{2\pi fC_p} = \frac{1}{r2\pi fC_p} \doteq \text{Power factor}$$

In the above formulas C_s and C_p are substantially equal for dissipation factors less than 0.1.

The power loss, P, in a parallel resistance is

$$P = \frac{E^2}{r} = E^2 2\pi fC_pD \text{ watts}$$

The power loss in a small series resistance is

$$P = E^2\,[2\pi fC_s]^2\,R = E^2 2\pi fC_sD \text{ watts}$$

Writing $C_s = C_p = C$ in terms of K, area in square centimeters A, thickness in centimeters t, and the dielectric coefficient for free space 8.854×10^{-12},

$$C = 8.854 \times 10^{-14} \cdot K \cdot \frac{A}{t}$$

and

$$P = 0.555 \times 10^{-12} \cdot E^2 \cdot \frac{AK}{t} \cdot f \cdot D \text{ watts}$$

which may be rewritten in terms of volume in cubic centimeters, V,

$$P = 0.555 \times 10^{-12} \left(\frac{E}{t}\right)^2 \cdot V \cdot f \cdot (K \cdot D) \text{ watts per cm}^3 \text{ per cycle per (volt per cm)}^2$$

Thus the power loss is seen to (1) increase as the square of the voltage gradient (E/t); (2) increase with the volume of material in the field; (3) increase with frequency if $(K \cdot D)$ is constant; (4) increase with the product $(K \cdot D)$, the loss factor. Shunt resistance or conductance is often the most important cause of losses at frequencies below 10 kc, since the impedance of the capacitor is relatively high compared to the shunt resistance. At a frequency of 100 kc the impedance will be much lower compared to the shunt resistance, and at 1000 kc the shunt resistance of good dielectric materials is extremely high compared to the impedance. In the frequency range from 1000 kc to 1000 Mc the dielectric constant and dissipation factor of good non-polar dielectric materials vary but slightly, thus indicating a constant loss per cycle.

Polar and Non-polar Materials. A polar substance is characterized by a permanent unbalance in the electric charges within a molecule. This unbalanced charge system is known as a "dipole" and tends to turn in an electric field. In liquids and soft solids which are polar, there is a free rotation of the dipoles at certain temperatures and applied frequencies, causing a very high loss.

Non-polar materials have no permanent charge unbalance: though the molecule may be distorted by an applied electric field, no tendency to rotate exists. Non-polar substances therefore are free of sharp loss peaks as the temperature or frequency is varied, any changes in dielectric constant and power factor occurring gradually. Whether a substance is polar or non-polar can usually be predicted from its chemical structure. Most hydrocarbons are non-polar and hence are numbered among the best dielectrics, e.g., polyethylene, polystyrene, mineral wax, and oil.

Dielectric Absorption in Solids. A pure capacitance may be completely charged or discharged almost instantaneously if the resistance in the external circuit is small. When the capacitor contains an imperfect dielectric, current continues to flow and the charge increases for a considerable period. The current slowly approaches a final value fixed by the insulation resistance of the dielectric. For this reason, insulation resistance readings are taken after the voltage has been applied for a standard time of 1 minute.

When the capacitance is discharged through a small resistance, a portion of the total charge will be instantaneously dissipated; this has been called the "free charge." If the circuit is opened, the capacitor will be found to have another and smaller charge known as the "residual" or "bound" charge. Sometimes this process may be repeated two or three times.

When an absorbing dielectric is subjected to an alternating electric field, the maximum charge is greater than the free charge but less than the total charge in a d-c field. The measured value of the dielectric constant decreases with increasing frequency, approaching a value corresponding to the "free charge." The instantaneous charge is not in phase with the applied a-c voltage, and a loss of energy results, heating the dielectric. The maximum loss occurs at a frequency which is equal to the reciprocal of the time constant of the charge or discharge current-time curve. When the frequency is increased until the time constant for one cycle is very short compared to the charging time constant with direct current, practically no loss remains.

The exact mechanism of absorption loss is not known, although many theories have been suggested, such as surface charges in non-homogeneous dielectrics (Maxwell and Wagner), space charges (Whitehead and Joffé), and dipoles (Debye).

Insulation Resistance. Insulation resistance is the ratio of the applied d-c voltage to the resultant current flowing, after 1 minute of voltage application, between two electrodes embedded in, or making contact with, the dielectric. The nature of the specimen determines whether the value represents principally surface resistance or volume resistance: when thin specimens with studs or bolts as electrodes are measured at high humidities, the value is more nearly representative of surface resistance; when electrodes of large area are applied to the faces of a slab and measured at low humidities the result is more nearly representative of volume resistance.

Surface Resistivity. Surface resistivity is the resistance between opposite edges of a square. Since the resistance of the body of a material is always in parallel with the surface resistance, the latter is measurable only when the volume resistance is much greater than the surface resistance, e.g., under conditions of high humidity and large ratio of surface to volume in the electric field.

The degree of surface contamination is an important factor since perspiration or other surface contaminant dissolves in the condensed moisture layer and increases its conductivity. While a layer of pure water 0.1 micron thick would result in a surface resistivity of about 3×10^{10} ohms at room temperature (a value commensurate with most wetted insulating materials at humidities above 90 per cent), a mere trace of salt on the surface would reduce the resistance by a factor of 10^{-3}.

The temperature coefficient of surface resistivity is negative.

Volume Resistivity. Volume resistivity is the resistance between opposite faces of a centimeter cube after the surface leakage is eliminated; it is expressed in ohm-centimeters. Volume resistivity is calculated from the resistance between two electrodes, one of which is completely surrounded by a guard electrode maintained at the same potential. The current due to the surface resistance flows through the guard circuit and does not influence the value of current in the guarded electrode circuit.

Volume resistivity has a negative temperature coefficient and often is found to have a negative voltage coefficient. Some materials, especially those of a fibrous nature, exhibit changes in resistance with polarity and the resistivity may change with time, owing to polarization or to water migration. This is known as the "Evershed effect."

Dielectric Strength of Solids. The dielectric strength of a material is the maximum potential that unit thickness can withstand without breakdown. The value obtained will depend upon sample thickness, temperature, applied frequency, wave form, electrode form, area and heat conductivity, the surrounding medium, and the rate and total time of voltage application.

In order that test values may be comparable, the American Society for Testing Materials has standardized these variables for specific classes of materials. The dielectric strength values for a given class of material, e.g., molded thermosetting plastics, are therefore comparable but do not bear a direct relation to the values for a different class of material, e.g., mica or oil.

The general effect of increase in thickness is to raise the total breakdown voltage: the increase for solid dielectrics is nearly linear for small thicknesses at room temperature but very much less than linear for large thicknesses or higher temperatures. Elevation of temperature invariably reduces the dielectric strength.

The peak value of the 60-cycle a-c breakdown voltage usually is less than that of the d-c breakdown voltage. The breakdown voltage decreases with increasing frequency, the rate depending on the loss factor of the material. Dielectric strength at 1 megacycle may be as low as 25 per cent of the 60-cycle value.

When an electrode is sharply curved, the potential gradient at the surface is raised according to the laws of electrostatics, and a reduction in breakdown voltage results. If the electrode area is increased, the probability of a weak dielectric spot within the electrode area is likewise increased, and the average dielectric strength is reduced.

The effect of time on breakdown voltage is best stated by Peek's equation:

$$g = g_0(1 - at^{-\frac{1}{4}})$$

where g = dielectric strength at time t seconds.
g_0 = dielectric strength at infinite time.
a = a constant.

Both g_0 and a vary with temperature and thickness. The formula is unsatisfactory for times less than 0.01 sec or for very long periods.

The general effect of placing dielectrics in series is to decrease the a-c breakdown voltage, since the voltage divides in inverse proportion to the dielectric constants if the resistivities are high. Laminated structures, unless bonded with a medium of the same dielectric constant or impregnated throughout with some medium, tend to have a lower dielectric strength than that of an equivalent thickness of homogeneous material. The presence of air or gas between laminae causes a pronounced reduction in dielectric strength.

Flashover. Flashover is an insulation failure by discharge between the electrodes over the surface of an insulator. Sometimes the insulator is permanently damaged by the flashover. In a uniform electric field the flashover voltage at low relative humidities approaches the dielectric strength of air as a limit. Increase in humidity causes a surface moisture film to form on the insulator, reducing the flashover voltage. Substances which are wetted by water form a more or less continuous film, and the flashover values are somewhat erratic, falling rapidly up to about 50 per cent relative humidity and then more slowly. Non-wetted substances such as waxes, on which the moisture condenses in droplets, show an almost linear, and quite consistent, decrease in flashover with increase in humidity. Higher temperatures and lower pressures both reduce the density of the air: therefore both factors decrease the flashover value. The flashover voltage at high humidities may be from three-quarters to one-half of the sparkover voltage in the absence of the insulator.

Arc Resistance. The power arc following a flashover or the breaking of contacts over the insulator surface subjects the surface to extreme heat, to chemical action, and to deposition of electrode material. Where exposure of the insulator to arcs cannot be avoided, it is important to know the degree of resistance to be expected. Glass, mica, and ceramic materials are quite resistant and become permanently conductive only by deposit of electrode material. Organic materials fail by carbonizing and become permanently conductive even with intermittent arcing. Unfortunately, the phenolic materials are very poor in this respect, failing in a few seconds when tested according to ASTM method D495. Plastics, such as polystyrene, which liberate volatile monomers, tend to blow the arc from the surface and fail in about 60 sec. Shellac, hard rubber, and vulcanized fiber are moderately arc resistant. Vinyl plastics and methacrylates likewise are somewhat resistant. The melamine resins are outstanding as a class: glass cloth melamine laminates last over 180 sec. Cellulose acetates and ethyl cellulose may be formulated to exceed 180 sec resistance. Non-refractory cold-molded compounds and glass-bound mica may last from 180 to 400 sec. Refractory cold-molded compounds last even longer, approaching the ceramics. Arc-resistant varnishes are available which considerably improve the rating of phenolic materials.

Test Methods. Methods of testing electrical insulating materials have been standardized by the American Society for Testing Materials (1916 Race St., Philadelphia 3, Pa.). A few of the more important standards are listed below:

Dielectric Constant and Power Factor	ASTM D150
Dielectric Strength	ASTM D149
Insulation Resistance	ASTM D257
Arc Resistance	ASTM D495

Reference should be made to the current *Index to ASTM Standards* for further information.

5. SOLID DIELECTRIC MATERIALS

Table 1 lists the important physical and electrical properties of solid materials useful for electrical insulation; it is followed by a list of trade names and additional information. The properties given are not intended to be design values but to illustrate the range of properties to be expected in a given class of material or a value typical of the class.

Table 1. Properties of Solid

Physical Properties at 25° C

Density, gm/cm³	Tens. Str., lb/in.² ×10³	Comp. Str., lb/in.² ×10³	Mod. of Elasticity ×10⁶	Flex. Str., lb/in.² ×10³	Coef. Lin. Ther. Exp. per °C ×10⁻⁶	Ther. Cond. per °C ×10⁻⁴ (Note A)	Max. Oper. Temp., °C (Note B)	% H_2O Absorption in 24 Hr	Material (See text also) (Note C)
1.3–1.5	5–6	23.0	0.25–0.4	9.0	90–100	5.0	60–80	0.3–0.4	Allyl resin, cast
2.6	8.5	75	15	20	6.9	60	1000	0.00–0.08	Alsimag * No. 35
2.7	10	85	5	22	7.3	60	1000	0.00–0.05	" No. 196
2.7	7.5	65	...	18	7.5	60	1000	0.00–0.08	" No. 211
1.05	Melt. pt. 250–325° C			3	44	...	180?	Low	Amber
1.21	10.0	20–22	0.65	15–20	54	...	107	0.08	Aniline formaldehyde resin
1.5	9–11	38–44	1.2	12–16	30	...	148	0.2–0.6	Same, glass mat
0.5–0.8	0.3–0.4	...	...	...	...	3.45	260	Very high	Asbestos paper (dry)
1.01–1.09	Melt. pt. 57–87° C		...	...	...	14–20	...	...	Asphalt (native)
1–1.17	Melt. pt. 27–122° C; flash pt. 200–320° C					...	...	...	" petroleum
0.9–1.07	Melt. pt. 27–162° C; flash pt. 175–290° C					...	...	...	" blown
1.2–1.3	5–11	26–33	0.7–1	12–17	50–110	3–4	125	0.05–0.07	Bakelite * resin, pure
0.96	Melt. pt. 60.5–63.5° C			...	...	...	...	...	Beeswax, yellow
1.32–1.39	4.5–10	5.1–27	0.5–0.57	10–18	80	...	...	7–14	Casein plastic
1.35–1.6	4–10	25	0.2–0.39	9	120–160	3.1–5.1	45	2.3–3.2	Celluloid,* clear
...	...	...	...	...	...	...	...	...	Cellulose, dry
1.28–1.32	7–12	...	...	...	...	...	...	...	" acetate, film
1.25–1.37	4–8	10–30	0.15–0.3	3.5–10	120–160	5.4–6.5	50–70	2.0–7.5	" " transparent sheet
1.25–1.56	2.5–9	11	0.26	...	100	5.3–8.7	50–70	2.0–4.0	" " pigmented sheet
1.25–1.4	3–11	11–27	0.1–0.4	5–13	80–160	5.0–9.0	60–70	2.0–7?	" " molded, gen. purpose
1.14–1.22	2.7–8.0	7.5–22.0	0.1–0.35	2.0–13	110–160	4.5–7.8	60–75	1.6–2.6	Cellulose acetate, butyrate, molded
1.17–1.22	2.8–6.0	...	...	4.8–10	120–190	4–8	60–75	1.0–1.7	" propionate
0.91–0.92	Melt. pt. 65–75° C		...	...	...	...	...	Nil	Ceresin wax
1.07	...	...	...	13	...	...	100	0.3	Cerex *
2.0–2.4	2–4	40–50	7	8–10	1.1–2.5	30	1200	0.05–5.0	Cordierite ceramic
...	...	...	...	...	...	...	...	...	Electrose,* black
2.5–3.8	...	...	...	...	10	20–30	300	Very low	Enamel, vitreous
1.05–1.2	0.4–3.0	...	>0.2	...	...	...	55	1.5–3.0	Ethyl cellulose, non-rigid
1.01–1.18	2.7–8.0	8–20	0.1–0.4	4–12	100–140	5.6	50–90	1.0–2.0	" " rigid, gen. pur.
1.1–1.3	0.5–3.0	...	0.1–0.2	...	120–180	3.8–6.2	35–60	1.3–2.0	" " cast
>1.3	>6.0	>30	...	>13	27	4–6	125	<60	Fiber, bone, vulcanized
>1.05	>6.0	>20	...	>12	27	4–6	125	<65	" commercial
2.3–2.9	2–5	10–30	8.5–13	...	7.5–11	17–25	...	Nil	Glass, crown (lime)
2.9–5.9	3–6	6–10	8.7–13	...	8–10.5	14–20	...	Nil	" flint (lead)
...	...	...	...	...	8.9	17	...	Nil	" commercial plate
2.25	...	40	8.86	...	3.2	27	...	Nil	" Pyrex,* chem. resistant
2.23	...	...	...	10	3.3	24	500	Nil	" " elec. #774
2.10	...	...	...	7	3.2	...	450	<0.01	" Multiform * #707
2.18	...	...	...	...	0.78	...	800	Nil	" Vycor * 790
3.44	5–8	22–40	8	15–20	8	13	300	0.03	Glass-bonded mica, gen. purpose
2.75–3.5	6–10	30–45	8	14–19	5–10	6–12	300–400	0.003–0.1	" " " low loss grade
3.8–3.9	>6	>20	8	13–19	6.4	8.3	300–400	0.06–0.1	" " " injection molded
...	0.57	...	...	...	...	...	200	...	Gummon *
0.96	Melt. pt. 190° C		...	...	200	4.8	...	...	Gutta percha
...	1.99	1.56	...	...	...	...	...	...	Hemit *
2.5–2.8	6.0	100	15.0	...	7.0	17.5	1000	Nil–0.1	Isolantite *
1.83–1.92	...	...	...	...	...	...	...	...	Ivory
2.5–2.7	...	20–30	...	...	...	19.2	1000	...	Lava *
2.8	2.0	20	...	8	8.3	50	1200	1.5	" (Grade I)
2.3	2.5	20	...	9	2.9	30	1100	2.5	" (Grade A)
2.66	7.2	96	...	10.5	8.1	...	...	0.01	Lavite *
1.3–1.4	7–11.5	24–29	0.7–1.4	15–19	21–24	...	70	0.05–4.2	Lignin, sheet
1.01–1.09	Nil	...	...	...	...	ca. 15	ca. 1000	ca. 60	Magnesium oxide, comp. powd., dry
...	...	...	...	...	...	...	...	...	Marble, blue
2.6–2.84	...	8.5–21.3	...	...	10–16	71–91	...	0.26	" white
1.35–1.4	7.8	...	...	12–13	...	...	...	0.4–1.9	Masonite die stock *
1.45–1.55	6–7.5	22–25	...	9–14	...	...	98	0.6–1.8	Melamine, alphacellulose filled
1.4–1.5	3.6–7	23	...	10.5–14	...	...	98	<1.5	" chopped rag filled
1.9–1.95	16–29	38–60	...	31–62	...	...	...	1.5–3.0	" glass cloth filled
1.7–2.2	5–6	20	1.6	7.5–9.3	20–45	...	125	0.1	" mineral filled

* Trademark names.

Dielectric Materials

Electrical Properties at 25° C

Dielectric Constant (Note D)		Power Factor % (Note D)		Volume Resistivity (Note E)		Surface Resistivity, ohms at 20–25° C			Dielectric Strength (Note F)		Reference, See page 2-32
						Relative Humidity					
Freq. less than 2 kc	300 to 2000 kc	Freq. less than 2 kc	300 to 2000 kc	Ohm-cm at 20–25° C	Temp. Coef. 20–30° C	30%	50%	90%	Thickness, mils	Volts per mil	
3.75–4.0	3.5	1.5	5.6	10^{11}–10^{14}	...	...	...	...	125	450	28, 33, 39
6.5	6.2	0.3	0.2	$>10^{14}$	...	...	...	...	250	225	39
6.3	6.0	0.14	0.08	$>10^{14}$	...	...	...	...	250	240	39
...	5.8	...	0.02	$>10^{14}$	...	...	...	...	250	240	39
2.86	...	...	0.513	5×10^{16}	...	...	6×10^{16}	3×10^{12}	...	1400	1, 16
3.73	3.6	0.23	0.66	10^{12}	...	...	...	10^{11}–10^{12}	125	600	39
4.6	4.5–5.2	2.0	1.0–1.3	10^{13}	...	...	...	3.1×10^{9}	62	450–600	32, 39
2.7	...	...	...	...	...	...	...	...	47	100	2, 7
2.7	...	...	...	...	...	...	...	...	90	25–50	8, 12, 17
3.1	...	2.29	...	6.1×10^{14}	...	...	...	...	...	...	8, 12, 17
3.1	...	2.29	...	6.1×10^{14}	...	...	...	...	...	...	8, 12, 17
...	4.5	...	0.2	2×10^{16}	2.6	4×10^{16}	8×10^{15}	8×10^{14}	...	250–700	1, 12, 26
2.88	3.2	2.94	1.63	2×10^{15}	16.0	7×10^{14}	6×10^{14}	5×10^{14}	...	250	1, 5, 8
...	6.15	...	5.19	1.1×10^{10}	...	...	5×10^{9}	...	125	160–700	26, 33
6.7–7.3	6.8	6.2–14.4	7.4–9.7	2×10^{10}	1.8	8×10^{10}	5×10^{10}	2×10^{9}	125	300–700	1, 12, 33
3.9–7.5	...	...	...	1×10^{9}	...	...	...	...	...	...	12
4.3–4.8	3.3–4.0	2.0–3.0	3.0–4.2	10^{14}	...	...	...	...	2–7	2800–3300	39
4.5–6	3.3–5	3–8	3–6	5×10^{10}	...	...	3×10^{12}	10^{11}	10–30	600–1300	12, 28, 39
12	4.9	15.3	...	...	...	...	...	...	...	600–1000	39
4.5–6.2	4.0–5.2	1.5–4	4–5.5	10^{11}–10^{13}	...	...	10^{10}–10^{12}	...	125	290–365	40
3.6–6.4	3.0–6.2	1–6	1–5.0	10^{9}–10^{12}	...	...	...	$>10^{13}$	125	250–400	32, 40
3.6–3.8	3.3–3.5	0.4–1.4	1.9–3.2	...	...	...	...	...	...	370–425	39
2.2	2.5	0.03	0.04	$>5\times10^{18}$	...	$>8\times10^{16}$	$>8\times10^{16}$	$>8\times10^{16}$	...	...	1, 5, 8
2.7	2.7	0.24	0.24	...	...	...	...	...	125	500	33
...	5–6	...	0.4–1.7	10^{13}–10^{14}	...	...	...	...	250	100	19, 21, 39
...	...	...	...	1×10^{14}	2.3	3×10^{12}	1×10^{12}	8×10^{9}	125	600	1, 12
...	...	...	...	...	...	...	...	...	...	...	16
2.6–3.0	...	0.25–0.6	...	10^{14}	...	...	...	...	...	...	40
2.5–4.0	3–3.7	0.5–2.5	0.7–4.0	10^{12}–10^{15}	...	...	...	10^{11}	125	250–600	32, 40
3.3–3.4	3.3–3.6	0.74–2.5	1.3–2.6	...	...	...	...	...	...	...	31
2.5	5–7.5	...	5.0	2×10^{10}	3.2	3×10^{10}	5×10^{9}	3×10^{7}	<60	>175	1, 5, 12, 41
2.5	5–7.5	...	5.0	2×10^{10}	3.2	3×10^{10}	5×10^{9}	3×10^{7}	>40	>175	1, 5, 12, 41
6–8	...	1.0	1.0	$>10^{13}$	...	...	...	...	...	1200	15, 23, 27
7–9	7	0.45	0.42	$>10^{13}$	...	...	...	...	...	860	15, 23, 27
...	7.6	...	0.82	2×10^{13}	3.2	8×10^{13}	5×10^{10}	2×10^{6}	...	...	1, 16
4.8	4.9	0.2–0.4	0.42	10^{14}	...	...	...	$>10^{13}$	...	...	15
...	4.7	...	0.3	5×10^{14}	...	...	...	...	...	1300	15
...	4.0	...	0.08	...	...	...	...	...	...	>500	27, 39
...	4.0	...	0.05	3×10^{13}	...	...	...	...	...	>500	27, 39
...	8.3	...	1.3	10^{15}	...	...	...	...	250	>180	39
...	6.5–7.5	...	0.15–0.2	10^{15}	...	...	10^{5}–10^{7}	...	250	>180	39
...	7.8–8.3	...	0.15–0.2	10^{15}–10^{17}	...	...	...	...	...	...	39
...	...	...	...	3×10^{12}	1.4	5×10^{12}	2×10^{12}	3×10^{8}	...	75	1, 12
3.0–4.9	...	1.8	...	2.5×10^{15}	...	...	...	...	...	200–500	1, 12
...	...	...	...	1×10^{10}	1.2	3×10^{10}	1×10^{10}	5×10^{8}	...	50–75	1, 12
6.1	6.1	...	0.18	2.7×10^{14}	...	...	1×10^{15}	...	130	320	39
...	...	...	...	2×10^{8}	1.5	2×10^{10}	5×10^{9}	1×10^{9}	...	...	1, 16
...	...	...	...	2×10^{10}	...	6×10^{11}	1×10^{11}	1×10^{8}	...	75–250	1, 16
...	5.6	...	0.3	...	...	...	...	...	250	<100	39
...	5.3	...	1.0	...	...	...	...	...	250	<80	39
...	6.4	...	0.45	3.2×10^{11}	...	...	...	...	...	235	39
...	4.1	...	3.7–32	...	...	...	...	...	125	500–650	39
2.2	2.5–3.5	0.25–0.6	0.2	...	...	...	...	...	8–16	300–700	20
8.3	9.4	0.3–5.0	1.22	1×10^{9}	...	5×10^{11}	8×10^{9}	1×10^{7}	...	...	1, 5, 7, 12
...	9.3	...	0.52	1×10^{11}	...	8×10^{10}	3×10^{9}	2×10^{7}	...	...	1, 5, 7, 12
5.7–6.8	...	4.2–5.3	...	2×10^{6}	...	...	4×10^{5}	...	...	...	39
8–9.5	7–8.2	3.7–8	2.9–6	10^{11}	...	...	...	...	125	300	40
7.7	5.6–7.2	8	3.8–4.1	10^{12}	...	...	...	...	125	270	40
6.9–7.5	...	...	1.1–1.3	...	...	...	...	2×10^{8}	125	500	32, 40
8–10	5.8–6.7	11	2.8–6	10^{13}	...	...	...	...	125	300	40

Table 1. Properties of Solid

Physical Properties at 25° C

Density, gm/cm³	Tens. Str., lb/in.² ×10³	Comp. Str., lb/in.² ×10³	Mod. of Elasticity ×10⁶	Flex. Str., lb/in.² ×10³	Coef. Lin. Ther. Exp. per °C ×10⁻⁶	Ther. Cond. per °C ×10⁻⁴ (Note A)	Max. Oper. Temp., °C (Note B)	% H_2O Absorption in 24 Hr	Material (See text also) (Note C)
1.18–1.2	5.8–9	10–12.5	0.3–0.5	12–14	70–90	4–6	62	<0.4	Methacrylate, cast, regular
1.18–1.2	6.5	...	...	>12	...	...	75	<0.4	" " heat-resistant
<1.2	4–8	10–15	0.4–0.6	12–16	60–80	5–7	50–58	0.4–0.6	" molded, regular
<1.2	5–10	15–25	0.4	10–15	60–80	5–7	65?	0.4–0.6	" " heat-resistant
2.6–3.2	...	...	...	...	...	12	500	...	Mica (muscovite)
2.6–3.2	...	...	...	...	...	12	500	...	" U.S.A. clear
2.6–3.2	...	...	...	...	...	12	500	...	" India clear
2.6–3.2	...	...	...	...	...	12	500	...	" stained
2.6–3.2	...	...	...	...	...	12	1000	...	" (Phlogopite)
2.3–2.4	...	...	...	...	...	5–8	125	...	" reconstructed plate
...	...	...	...	...	...	5–8	...	...	" flexible
...	1.1	Elongation 650%		...	...	...	150	...	Neoprene GN
1.06–1.19	40–50	...	1–1.6	...	...	...	...	1.5?	Nylon * filament
1.06–1.19	5–9	18	0.3–0.45	12–15	100	...	120	1.5?	" molded
0.85–0.95	Melt. pt. 65–90° C		...	...	...	...	...	...	Ozokerite
0.8–1.0	...	...	...	...	...	6.4	...	...	Paper, kraft, dry
0.87–0.91	Melt. pt. 58° C		...	...	130–400	4.7–6.2	...	...	Paraffin
0.86–0.88	Melt. pt. 38–50° C		...	...	...	2.2–3.8	...	...	Petrolatum
1.3–1.36	9–15	35	0.4–2.0	16–24	17–25	5–8	100–110	4–6	Phenolic laminates, Grade X
1.3–1.36	6–10	22	0.4–2.0	11–19	17–25	5–8	100–110	3–5	" " " P
1.3–1.36	8–12	34	0.4–2.0	12–20	17–25	5–8	100–110	1.3–2	" " " XX
1.3–1.36	6–10	25	0.4–2.0	12–20	17–25	5–8	100–110	1.3–2	" " " XXP
1.3–1.36	6–8	32	0.4–2.0	12–18	17–25	5–8	100–110	1–1.2	" " " XXX
1.3–1.36	5–8	25	0.4–2.0	12–18	17–25	5–8	100–110	1–1.2	" " " XXXP
1.3–1.36	7.5–12	35–44	0.35–1.5	16–28	17–30	5–8	100–125	2–4.4	" " " C
1.3–1.36	6.5–10	34–38	0.35–1.5	13–21	17–30	5–8	100–125	1.2–1.8	" " " CE
1.3–1.36	7–11	30–38	0.35–1.5	15–28	17–30	5–8	100–125	1.5–2.5	" " " L
1.3–1.36	6.5–10	33–40	0.35–1.5	15–28	17–30	5–8	100–125	1.2–1.8	" " " LE
1.5–1.8	5–12	30–40	0.35–1.5	10–20	17–25	...	150–200	1.0–1.5	" " " A
1.5–1.8	7–18	35–44	0.35–1.5	16–30	17–25	...	150–200	1.2–1.9	" " " AA
1.4–1.6	14–19	42–47	1–1.2	20–25	...	...	...	0.3–0.5	" " glass fabric
1.2–1.3	6–9	25–30	0.7–1.0	11–17	25–60	8–12	125	0.05–0.2	" moldings, unfilled
1.3–1.45	6–8	20–30	1–1.5	9–14	30–75	4–7	135–148	0.5–1.6	" " wood flour, g.p.
1.35	6–10	20–30	1	11	30	5	135	1	" " organic elec.
1.40–1.45	6–7.5	25–30	0.7–1.2	9–11	30–60	4–7	115	1–1.75	" " chopped rag filled
1.6–2.0	4–10	15–30	1–1.5	8–20	25–40	8–20	150–200	0.05–0.2	Phenolic moldings, heat-resistant asbestos filled
1.8–2.0	>4.5	15–30	3.5?	8–12	...	...	135	0.007–0.07	Phenolic moldings, mica filled
1.27–1.32	5–12	10–30	0.3–1.0	9–15	28	3–5	70	0.01–0.5	" cast, unfilled
0.92	1.7–3.0	3	0.015	1.7	170–210	6.2–8.1	75–80	Nil–0.01	Polyethylene
1.38	...	...	...	...	...	...	82	0.05	Polydichlorostyrene
>1.05	5–9	11.5–15	0.16–0.47	>8–12	60–80	1.9	65	<0.1	Polystyrene, general purpose
1.05–1.07	>5	14	...	>8	60–80	1.9	65	<0.1	" best elec.
1.3–1.4	>3	...	...	>8	60–80	...	75	<0.1	" heat-resistant
2.1–2.3	2–4.5	1.7	0.06?	2	55	5.8	200	Nil	Polytetrafluoroethylene
2.3–2.5	5–6	45–60	7–15	10–11	3–6	25–50	1000	Nil	Porcelain, unglazed, wet proc.
2.2–2.5	1–2	30–50	...	6–8	3–4	25–50	1000	0.5–1.0	" " dry proc.
1.1	...	...	...	...	...	4.5	90	High	Pressboard, untreated
1.1	...	...	...	...	...	4.5	90	...	" oiled
1.1	...	...	...	...	...	4.5	90	...	" varnished
2.4	5	45	...	11	...	21	...	0.025	Prestite *
2.2	7.0	200.0	10.15	...	0.57	36	1000	Nil	Quartz, fused
1.07	Melt. pt. 70–100° C			...	...	...	...	...	Rosin
1.1–1.4	1.5–10	2–5	0.33	...	70–85	3.8–8.7	<65	0.03–0.08	Rubber, hard
0.91	0.13–0.35	...	...	...	670	...	...	2–4	" natural, unvulcanized
0.92	3.2–4.2	...	...	...	660	3–3.8	...	2–4	" " vulcanized
...	...	...	...	...	...	...	...	...	" " 60% zinc oxide

* Trademark names.

Dielectric Materials—*Continued*

Electrical Properties at 25° C

Dielectric Constant (Note D)		Power Factor % (Note D)		Volume Resistivity (Note E)		Surface Resistivity, ohms at 20–25° C			Dielectric Strength (Note F)		Reference, See page 2-32
						Relative Humidity					
Freq. less than 2 kc	300 to 2000 kc	Freq. less than 2 kc	300 to 2000 kc	Ohm-cm at 20–25° C	Temp. Coef. 20–30° C	30%	50%	90%	Thickness, mils	Volts per mil	
<4	<3	6–7	1.5–4.0	$>10^{14}$	......	......	....	$>10^{12}$	125	500	40
<4	<3	<7	<4	$>10^{14}$	......	......	....	....	......	....	40
<4	<3	3.8–7	1.5–4	$>10^{14}$	......	......	....	....	125	500	40
<4	<3	7	<4	$>10^{14}$	......	......	....	....	......	....	40
4.5–7.5	....	0.1–7.0	....	1.3×10^{14}–2×10^{17}	2.0	1×10^{14}	2×10^{13}	8×10^{9}	......	....	1, 7, 12, 16
....	6.57–8.59	....	0.01–0.04	....	......	......	....	....	1–11	1500–5700	14
....	7.07–7.90	....	0.01–0.02	....	......	......	....	....	1–11	1300–4200	14
....	5.83–9.64	....	0.06–8.36	....	2.7	1.5×10^{12}	1×10^{10}	9×10^{7}	1–11	1300–4100	1, 14
....	5.41–6.07	....	0.38–7.12	4.5×10^{11}–2×10^{13}	......	......	....	....	1–11	1500–5000	12, 14
....	3.4–4.1	....	0.13–0.32	2×10^{11}	......	......	....	....	6	950	39
....	....	....	....	....	......	......	....	....	6	600	39
8.3	....	1.6	....	4.2×10^{12}	......	......	....	....	......	....	40
4.5	....	2.7	....	10^{14}	......	......	....	....	......	....	39
3.2–4.5	3.3–4	1.0–2	2.2–2.5	10^{13}	......	......	....	2×10^{10}	125	400	32, 40
2.4	....	0.92	....	5×10^{14}	......	......	....	....	25	1100	12
3.5	....	0.5	....	....	......	......	....	....	6	750–1000	12
2.2	2.2	0.03	0.02	$>5\times10^{18}$	......	$>1\times10^{18}$	....	$>1\times10^{18}$	......	....	1, 5, 12
2.2	....	0.29–0.5	....	2×10^{12}–10^{13}	......	......	....	....	100	500	10, 12
....	....	....	....	....	......	......	....	....	62	500–700	40, 41
....	....	....	....	....	......	......	....	....	62	500–700	40, 41
....	4.7–5.5	....	3.8–4.5	10^{10}–10^{13}	......	......	....	5.6×10^{9}	62	500–700	40, 41
....	4.7–5.5	....	3.8–4.5	....	......	......	....	....	62	500–700	40, 41
....	4.5–5.2	5.0	3.0–3.5	....	......	......	....	....	62	500–650	40, 41
....	4.2–5.2	....	2.4–3.0	....	......	......	....	3×10^{9}	62	500–650	40, 41
....	7	....	10	10^{9}–10^{12}	......	......	....	9×10^{8}	62	200	40, 41
....	5–6	....	4.5–6.5	10^{9}–10^{12}	......	......	....	....	62	400–500	40, 41
....	7	....	10	10^{9}–10^{12}	......	......	....	....	62	200	40, 41
....	4.5–5.5	12.0	3.5–5.5	10^{9}–10^{12}	......	......	....	1.8×10^{9}	62	400–500	40, 41
....	....	....	....	....	......	......	....	....	125	110–225	40, 41
....	....	7.5	15	....	......	......	....	....	125	50–150	40, 41
....	3.7–4.1	....	0.9–1.3	....	......	......	....	....	62	500–600	39
5–7	4.5–6	5–10	1.5–3.0	10^{12}	......	......	....	....	125	300	40
5–12	4.5–8	4–35	3.5–9	10^{8}	......	......	....	....	125	200–350	40
6.25	5.5	7.3	4.3	10^{11}	......	......	....	....	125	550	40
5–10	4.5–6.5	8–20	5–10	10^{9}–10^{11}	......	......	....	2.4×10^{8}	125	400	40
5–20	4.5–20	10–18	4–10	10^{9}–10^{11}	......	......	....	1.6×10^{10}	125	300–600	40
6 max.	4.5–5	1.0–2.5	0.7–1.5	10^{12} max.	......	......	....	10^{8}–10^{10}	125	550	40
5–10	5–7	2.5–20	1.0–4.5	10^{9}–10^{14}	......	......	....	....	125	350–450	40
2.3	2.3	0.03–0.05	0.03–0.05	10^{15}–10^{17}	......	10^{14}	....	1.7×10^{10}	62	1000–1100	39
....	2.6	....	0.02	....	......	......	....	10^{10}–10^{12}	......	....	33
2.5–2.75	2.5–2.75	0.01–0.05	0.02–0.1	10^{15}	......	......	....	2.4×10^{10}	125	>450	40
2.5–2.7	2.5–2.7	0.01–0.04	0.016–0.04	10^{17}–10^{19}	......	......	$>10^{17}$	$>10^{13}$	125	500–700	40
2.6–2.8	2.6–2.8	0.04–0.3	0.02–0.1	3.5×10^{19}	......	......	....	....	125	400	40
2.0	2.0	<0.02	<0.02	10^{16}	......	......	....	$>10^{13}$	80	480	39
6–7	6–7	1.0–2.5	0.8–1.0	10^{14}	1.6	2×10^{13}	6×10^{11}	5×10^{6}	250	250	1, 12, 13
6–7.5	6–7.5	1.7–2.5	0.8–1.0	10^{8}–10^{12}	......	......	....	....	250	40–100	19, 39
2.9–4.5	....	....	....	1×10^{9}	......	......	....	....	80–120	125–300	12
5.0	....	....	....	....	......	......	....	....	60	750	11
3.0	....	....	....	....	......	......	....	....	60	400	11
7.65	6.08	2.76	0.9	6×10^{11}	......	......	....	....	......	....	39
....	4.2	....	0.02	$>5\times10^{18}$	......	1×10^{15}	3×10^{12}	2×10^{8}	500	200	1, 9, 13
2.73	3.3–4.7	0.287	0.25–0.4	5×10^{15}–5×10^{16}	3.6	8×10^{14}	5×10^{14}	2×10^{14}	......	....	1, 8
2.8–3.5	3	0.4–0.5	0.6–2.1	10^{18}	......	6×10^{15}	3×10^{15}	2×10^{9}	80	250–900	1, 4, 5, 12
2.3–2.5	2.3–2.4	0.1–0.3	0.1–0.2	10^{15}–10^{16}	......	......	....	....	......	....	6
2.4–2.9	2.4–2.7	0.4	0.4	10^{15}–10^{16}	......	......	....	....	......	500–700	6
5.01	....	0.81	....	3.5×10^{15}	......	......	....	....	......	....	6

Table 1. Properties of Solid

Physical Properties at 25° C

Density, gm/cm³	Tens. Str., lb/in.² ×10³	Comp. Str., lb/in.² ×10³	Mod. of Elasticity ×10⁶	Flex. Str., lb/in.² ×10³	Coef. Lin. Ther. Exp. per °C ×10⁻⁶	Ther. Cond. per °C ×10⁻⁴ (Note A)	Max. Oper. Temp., °C (Note B)	% H_2O Absorption in 24 Hr	Material (See text also) (Note C)
...	...	...	...	...	...	...	...	...	Rubber, natural, 20% carbon black
...	>0.5	Elongation >200%			...	...	50	...	" wire insulation, Code
...	>1.2	Elongation >400%			...	...	60	...	" " " Performance
...	>1.5	Elongation >400%			...	...	75	...	Rubber, wire insulation, heat-resistant
...	>1.5	Elongation >450%			...	...	70	...	Rubber, wire insulation, low water absorption
...	>1.5	Elongation >450%			...	...	70	...	Rubber, wire insulation, low power factor
0.94–0.98	1.5–3.2	Elongation 400–650%			...	...	120–140	...	Rubber, synthetic, Buna S, gum stock
0.96–1.03	2–4.5	Elongation 400–800%			...	...	120–140	...	Rubber, synthetic, Buna N, gum stock
0.91	3.2	Elongation 800%			...	...	145	...	Rubber, synthetic, butyl, gum stock
0.91	0.3–1.5	Elongation 600–1000%			...	...	60	...	Rubber, synthetic, polyisobutylene, gum stock
1.24	1.6–1.75	Elongation 400–435%			...	...	145	...	Rubber, synthetic, Neoprene, gum stock
1.34–1.6	0.8–1.4	Elongation 350–600%			...	...	90	...	Rubber, synthetic, polysulfide, gum stock
0.97	...	...	...	...	...	...	<50	<0.1	Rubber, synthetic, cyclicized natural rubber
1.65–1.75	4–6	7.5–10	0.04–0.24	15–17	160–190	2.2	70–90	00–0.1	Saran,* molded
1.009	Melt. pt. 45–75° C		...	...	...	6.0	...	...	Shellac
1.1–2.7	0.9–2.0	7	...	3.0	...	...	40–60	<0.25	" compound
...	...	...	...	28–30	...	...	150–200	<0.3	Silicone glass laminate
...	...	...	...	...	...	...	175	...	" varnish
...	...	...	...	...	...	...	175	...	" varnished glass cloth
0.98–1.00	...	...	...	...	...	...	200?	...	" sealing compound
0.968–0.973	...	...	...	...	...	...	...	...	" fluid
1.5–2.0	0.20–0.33	...	...	...	...	...	...	...	" rubber (Silastic *)
2.6–3.3	3.5–10	10–14.2	8	>8	10.5–20	48	120–200	0.2	Slate
2.5–2.6	6.5–10	65–90	13–15	18–24	6.5–8.5	60	980–1000	Nil–0.1	Steatite ceramics, regular
2.5–2.8	6.5–10	65–90	13–15	18–24	6.5–8.5	60	980–1000	Nil–0.1	" " low loss
0.95–0.97	0.9–1.2	...	...	...	180–220	4.32	60–90	0.2–0.5	Styraloy * 22
1.36	3.3	11	0.33	6.5	...	...	80	0.05	Styramic *
1.38	...	...	...	...	...	...	110	0.03	" HT
2.0–2.1	Melt. pt. 112.8° C		...	...	64	7	95	...	Sulfur
...	1.2	1.14	...	...	...	...	200	...	Tegit *
									Titanium ceramics:
3.9–4.05	5–7.5	80–100	15	18–20	7–8	...	300	Nil	Titanium dioxide
...	...	...	...	...	...	...	...	...	Barium titanate
4	...	...	...	...	...	...	...	...	Calcium titanate
3–3.6	4–5	40–80	...	16–19	5–7	...	300	Nil	Magnesium titanate
...	...	...	...	...	...	...	...	...	Strontium titanate
1.45–1.6	6–10	24–30	1.2–1.5	>10	25–30	7	77	1–3	Urea formaldehyde, cellulose filled
<1.55	>6	>24	...	>9	...	...	77	1–3	" " arc resistant
...	...	...	...	...	...	...	...	...	Varnish, insulating
1.24	Tens. str. 40 lb. per in. width				...	5.0	85	...	Varnished cloth, yellow
1.26	Tens. str. 40 lb. per in. width				...	6.0	90	...	" " black
									Vinyl plastics:
1.26	2–6	...	...	...	...	5.0	100	Attacked	Polyvinyl alcohol
1.05–1.5	1–7	...	...	...	150	...	60	<3.0	" butyral
...	...	...	...	...	...	...	...	...	" carbazole
...	3–14	...	...	...	65	...	70–100	1.3	" formal
1.22–1.65	1–9	...		...	...	ca. 4.0	60–80	0.4–1.0	" chloride, filled, non-rigid
									Vinyl chloride-acetate:
1.32–1.36	>8	...	>0.38	>13	69	4.0	45	<0.15	Clear sheets
1.4–1.55	>6	...	>0.36	>11	...	...	47	<0.15	Sheets

* Trademark names.

Dielectric Materials—*Continued*

Electrical Properties at 25° C

Dielectric Constant (Note D)		Power Factor % (Note D)		Volume Resistivity (Note E)		Surface Resistivity, ohms at 20–25° C			Dielectric Strength (Note F)		Reference, See page 2-32
						Relative Humidity					
Freq. less than 2 kc	300 to 2000 kc	Freq. less than 2 kc	300 to 2000 kc	Ohm-cm at 20–25° C	Temp. Coef. 20–30° C	30%	50%	90%	Thickness, mils	Volts per mil	
5.97	...	8.8	...	3×10^{13}	...	...	...	...	...	...	6
4.5–6.0	...	5.0–7.0	...	$>5\times10^{14}$	2.4	...	...	...	...	375–425	28, 39
5.0–6.0	...	4.0–6.0	...	$>2.7\times10^{15}$	2.4	...	...	...	...	450–550	28, 39
5.0–6.0	...	4.0–6.0	...	$>2.0\times10^{15}$	2.4	...	...	...	...	450–550	28, 39
2.75–3.0	...	0.8–1.3	...	$>3.2\times10^{15}$	...	...	...	...	...	600–700	39
2.75–3.0	...	0.8–1.3	...	$>2.7\times10^{15}$	...	...	...	...	...	...	39
2.7–4.4	...	1.7	...	10^{15}–10^{16}	...	...	...	...	...	750	40
14–19	...	3.8–9.0	...	10^{9}–10^{11}	...	...	...	...	...	500	40
2.1	...	0.04	...	10^{18}	...	...	...	...	...	...	40
2.2–2.4	...	0.02–0.05	...	10^{16}–10^{18}	...	...	...	...	...	400–500	40
7.5	...	3	...	...	...	...	...	...	...	...	40
7.5	...	50	...	...	...	...	...	...	...	...	40
...	2.6–2.7	...	0.06–0.12	10^{16}	...	...	...	...	...	620	40
4–6	3–3.3	3–8	4.5–6.5	10^{14}–10^{15}	...	...	...	...	125	>350	28, 39
3–3.7	4.1	0.81	2.5	1×10^{16}	1.5	2×10^{14}	5×10^{13}	6×10^{9}	0.8	900	1, 7, 8, 12
...	...	...	...	6×10^{10}–2.3×10^{11}	...	...	1×10^{12}	...	200	200	40
...	...	...	<0.5	...	...	...	...	...	...	...	39
3–3.5	...	0.7	...	...	...	...	...	...	...	...	39
3–4	...	0.3–0.7	...	...	...	...	...	...	7	1500–2000	39
2.8	2.8	0.05–0.07	0.05	10^{13}–10^{14}	...	...	...	...	10	500	39
2.4–2.85	2.4–2.82	0.01	0.02–0.06	10^{14}	...	...	...	...	100	250–300	39
...	5–7.5	...	0.13–0.18	...	...	...	...	$>10^{13}$	...	500	39
6–7.5	<30	8.6	<63	1×10^{8}	...	2×10^{8}	9×10^{6}	1×10^{6}	1000	5–10	1, 5, 6, 12
5.5–6.5	5.5–6.5	0.13–0.3	0.1–0.2	10^{14}–10^{16}	...	...	...	$>10^{9}$	250	200–250	39
5.5–6.5	5.5–6.5	<0.15	0.04–0.1	10^{14}–10^{16}	...	...	...	$>6\times10^{8}$	250	200–250	39
2.5–2.6	2.4–2.6	0.07–0.12	0.05–0.15	10^{20}	...	...	...	...	125	700–800	39
2.55	2.5	0.11–0.26	0.04	...	...	...	...	...	...	800	40
...	2.6	...	0.02	...	...	...	...	10^{10}–10^{12}	...	...	33
3.6–4.22	3.8	...	...	1×10^{17}	4.9	1×10^{16}	7×10^{15}	1.1×10^{15}	...	...	1, 16
...	...	...	...	2×10^{12}	1.4	2×10^{12}	...	5×10^{7}	...	50	1, 12
...	80–100	...	0.04–0.07	10^{13}–10^{14}	...	...	...	...	250	100–200	37, 39
...	1200	...	<0.1	...	...	...	...	...	...	...	37
...	165	...	<0.1	...	...	...	...	...	...	...	37
...	14–18	...	0.007	$>10^{14}$	...	...	...	...	250	100–200	37
...	275	...	<0.1	...	...	...	...	...	...	...	37
7.5–9.5	6.6–8.2	4–6	2.7–4.6	10^{11}–10^{13}	...	10^{11}–10^{12}	...	...	125	300–720	40
7.5–9.5	<8.2	<10	<4.0	$>10^{10}$	...	...	...	...	125	300–720	40
...	4.8	...	5.12	...	...	8×10^{13}	1×10^{13}	1×10^{9}	1	700–1000	5
4.5–5.5	2.5	8	3.0	10^{9}	6	...	...	...	10	900–1200	2, 3, 11
4.5–5.5	2.0	6	2.0	...	...	...	...	...	10	800–1100	2, 3, 11
...	...	...	...	10^{7}	...	...	...	...	...	...	40
3.6–3.7	3.3–3.5	6	...	10^{14}	...	...	...	...	...	300–400	40
...	2.9	...	0.4	...	...	...	...	...	...	...	40
3.8	3.7	0.7	0.9	10^{15}	...	...	...	...	...	500–700	40
4–12	...	13.6	...	...	...	...	...	...	75	600–2000	40
3.2–3.5	3–3.3	<1.3	<1.9	$>10^{14}$	...	...	...	4×10^{10}	125	600	40
3.2–3.5	3–3.3	<1.3	<1.9	$>10^{14}$	...	...	...	...	...	400	40

Table 1. Properties of Solid

Physical Properties at 25° C

Density, gm/cm^3	Tens. Str., lb/in.^{2} $\times 10^3$	Comp. Str., lb/in.^{2} $\times 10^3$	Mod. of Elasticity $\times 10^6$	Flex. Str., lb/in.^{2} $\times 10^3$	Coef. Lin. Ther. Exp. per °C $\times 10^{-6}$	Ther. Cond. per °C $\times 10^{-4}$ (Note A)	Max. Oper. Temp., °C (Note B)	% H_2O Absorption in 24 Hr	Material (See text also) (Note C)
									Vinyl chloride-acetate (*Contd.*)
1.30–1.35	>7.5	9–12	0.35–0.4	12–13	69	4	45	<0.15	Transparent, molded........
1.3–1.4	>4	>9	....	7.5–12	....	......	45	<0.15	Molded. ....
1.3–1.4	>5	9–12	0.35–0.85	8–12	....	......	47	<0.15	Opaque molded ....
1.15–1.29	1–3	....	....	....	....	......	60–70	0.3–0.7	Non-rigid.... ....
1.30–1.45	1.7–3.0	....	....	....	....	......	60–70	0.5–0.9	Filled, non-rigid............
1.68–1.75	4–6	9–10	0.08–0.17	15–17	190	2.2	70–90	0.1	Vinylidene chloride..............
0.62–0.75	0.77	6–8.6	....	7.0	6.4	4.3–10.4	....	....	Wood, maple, hard.............
....	....			....	....	....	....	....	" " paraffined.........
0.69–0.96	0.77	6–7.2	....	13.0	4.9	5–8	....	....	" oak.....................
3.7	12.7	90	....	25	4.9	120	1000	Nil	Zircon porcelain................

Note A. Thermal conductivity is expressed in $10^{-4} \times$ gram-calories per square centimeter per second for a temperature gradient of 1 deg cent per centimeter. The values are typical for the temperature range 0 to 100 deg cent. The temperature gradient is perpendicular to the laminations of laminar materials.

Note B. Maximum operating temperatures given are based on satisfactory operation under average conditions without excessive cold flow, distortion, or shortening of the operating life of the material. In many cases it will be necessary to reduce the operating temperature in order to obtain electrical properties or to reduce cold flow at higher unit stresses. In some cases higher operating temperatures may be used, particularly where the material can be obtained in special grades for this purpose. If thermal shock is involved the limit for many materials, especially the ceramics, will be much reduced.

Note C. Most of the materials listed are intended to be representative of a class rather than of a single sample, and the ranges of values should be interpreted accordingly. Since complete mechanical and electrical data seldom are available for a single sample or batch, and since data have been assembled from many sources, it is necessary to exercise good judgment in comparing the various materials. Values for laminated phenolic materials are those for sheets; properties of tubes are somewhat poorer.

Note D. Where possible, the ranges of dielectric constant and power factor include the variations to be expected over the indicated frequency ranges, although most of the values up to 2 kc are for 60

REFERENCES

1. Curtis, H. L., Insulating Properties of Solid Dielectrics, *Sci. Papers Bur. Standards*, No. 234 (1915).
2. Taylor, T. S., The Thermal Conductivity of Insulating and Other Materials, *Elec. J.*, December 1919, pp. 526–532.
3. Flight, W. S., *J. I.E.E.*, Vol. 60, 218–235 (1922).
4. Hoch, E. T., Power Losses in Insulating Materials, *Bell Sys. Tech. J.*, November 1922.
5. Preston, J. L., and E. L. Hall, Radio Frequency Properties of Insulating Materials, *QST*, Vol. 9, 26–28 (February 1925).
6. Curtis, H. L., and A. T. McPherson, Dielectric Constant, Power Factor and Resistivity of Rubber and Gutta Percha, *Technologic Papers Bur. Standards*, Vol. 19, 669–722 (May 1925).
7. Monkhouse, A., *Electrical Insulating Materials.* Sir Isaac Pitman and Sons, London (1926).
8. Lee, J. A., and H. H. Lowry, *Ind. Eng. Chem.*, February 1927, pp. 387–395.
9. Sosman, R. B., *The Properties of Silica.* Chemical Catalog Co., New York (1923).
10. Shanklin, G. B., and G. M. J. Mackay, Progress in High Tension Underground Cable Research and Development, *Trans. Am. Inst. Elec. Eng.*, Vol. 48, 338–372 (1929).
11. Peek, F. W., Jr., *Dielectric Phenomena in High Voltage Engineering.* McGraw-Hill, New York (1929).
12. *International Critical Tables.* McGraw-Hill, New York (1929).
13. Barringer, L. E., Mycalex—A Moulding Material with Unique Properties, *Gen. Elec. Rev.*, July 1931, pp. 406–409.
14. Lewis, A. B., E. L. Hall and F. R. Caldwell, Some Electrical Properties of Foreign and Domestic Micas and the Effect of Elevated Temperature on Micas, *Bur. Standards J. Research*, August 1931, pp. 403–418.
15. Littleton, J. T., and G. W. Morey, *Electrical Properties of Glass.* John Wiley, New York (1933).
16. *Smithsonian Physical Tables* (1933).
17. Abraham, H., *Asphalts and Allied Substances.* Van Nostrand, New York (1938).
18. Hartshorn, L., Plastics as Insulators, *J. I.E.E.*, Vol. 38, 474 (1938).
19. Thurnauer, H., Ceramic Insulating Materials, *Elec. Eng.*, Vol. 59, 451 (November 1940).
20. Gemant, A., and F. A. Glassow, Compressed Magnesia as an Electrical Insulator, *J. Franklin Inst.*, Vol. 230, 481 (1940).

Dielectric Materials—*Concluded*

Electrical Properties at 25° C

Dielectric Constant (Note D)		Power Factor % (Note D)		Volume Resistivity (Note E)		Surface Resistivity, ohms at 20–25° C			Dielectric Strength (Note F)		Reference, See page 2-32
						Relative Humidity					
Freq. less than 2 kc	300 to 2000 kc	Freq. less than 2 kc	300 to 2000 kc	Ohm-cm at 20–25° C	Temp. Coef. 20–30° C	30%	50%	90%	Thickness, mils	Volts per mil	
3.2–3.5	3–4.0	1–4	1.8	$>10^{14}$	......	....	....	....	125	>600	40
....	....	....	....	$>10^{14}$	......	....	....	....	125	>650	40
4.7	4	4–4.8	1–3	$>10^{14}$	......	....	....	....	......	....	40
....	....	....	....	$>10^{8}$	......	....	....	....	75	>400	40
6–9.5	....	6–12	....	10^{4}–10^{8}	......	....	....	....	75	>400	40
3–5.1	3–4	3–8	3–6.5	$>10^{14}$	......	....	....	....	125	350–2000	40
....	4.4	....	3.33	....	......	....	....	....	......	....	2, 3, 16
4.1	....	....	....	3×10^{10}	3.6	1×10^{12}	8×10^{11}	2×10^{9}	600	110	1, 12
3.64–6.84	3.3	....	3.85	....	......	....	....	....	......	....	3, 16
....	9.2	....	0.13	ca. 10^{14}	......	....	....	....	250	240	36

cycles or 1 kc, and most of the values in the higher frequency ranges are for 1 Mc. For good insulating materials, the dielectric constant does not change greatly at higher frequencies, but the power factor may rise or fall considerably. An auxiliary table of power factors, following the main table, gives data for some substances.

Note E. Volume resistivity has a large negative temperature coefficient. The values shown for temperature coefficient are the ratio of the resistivity at 20 deg cent to the resistivity at 30 deg cent. It should be remembered that at higher operating temperatures the order of merit of any two materials may be reversed. Furthermore, the volume resistivity may be seriously reduced by prolonged exposure to high humidities.

Note F. Dielectric strength given is for the short-time test at 60 cycles except in the case of the ceramics, where it is given for the step-by-step method. Only those values for the same thickness are directly comparable. Where two thicknesses are given, the higher dielectric strength applies to the lower thickness. The dielectric strength will be much less at higher temperatures or for long times (see discussion of dielectric strength). In the case of laminated materials, the electric field is perpendicular to the laminations; dielectric strength with the field parallel to the laminations may be very low.

Note G. For all wood the tensile strength and flexural strength are given for forces perpendicular to the grain, the compressive strength for forces parallel to the grain. Under Thermal Conductivity, the first figure is perpendicular, the second parallel, to the grain.

21. Rosenthal, E., The Electrical Properties of High Frequency Ceramics, *Electronic Eng.*, September and October 1941.
22. Rosenthal, E., and J. E. Nickless, Ceramic High Frequency Insulators, *Wireless World*, November 1941.
23. Shand, E. B., The Dielectric Strength of Glass—an Engineering Viewpoint, *Elec. Eng.*, Vol. 60, 41 (March 1941).
24. Simons, H. R., and C. Ellis, *Handbook of Plastics.* Van Nostrand, New York (1943).
25. *Technical Data on Plastics Materials.* Plastics Materials Manufacturers' Association, Washington, D. C. (1943).
26. Scribner, G. K., *A Ready Reference for Plastics.* Boonton Molding Co., Boonton, N. J. (1944).
27. Guyer, E. M., Electrical Glass, *Proc. I.R.E.*, Vol. 32, 743 (December 1944).
28. *ASTM Standards*, Vol. III (1944) (specifications for various plastics).
29. Englund, C. R., Dielectric Constants and Power Factors at Centimeter Wave Lengths, *Bell Sys. Tech. J.*, Vol. 23, 114 (January 1944).
30. Works, C. N., T. W. Dakin and F. W. Boggs. A Resonant-Cavity Method for Measuring Dielectric Properties at Ultra High Frequencies, *Proc. I.R.E.*, Vol. 33, 245 (April 1945).
31. Barron, H., *Modern Plastics.* John Wiley, New York (1945).
32. Field, R. F., The Effect of Humidity on Plastics Insulation, *Plastics and Resins*, September 1945.
33. Brouse, H. L., A Review of Plastic Materials, *Proc. I.R.E.*, Vol. 33, 825 (December 1945).
34. Bass, S. L., and T. A. Kauppi, Silicones—a New Class of High Polymers of Interest to the Radio Industry, *Proc. I.R.E.*, Vol. 33, 441 (July 1945).
35. Plastics Design is Given Impetus by Engineering Classification (SPI), *Elec. Mfg.*, Vol. 36, 128 (November 1945).
36. Russell, R., and W. C. Mohr, Zircon Porcelain, New Insulation for New Products, *Elec. Mfg.*, January 1946.
37. Wainer, E., High Titania Dielectrics, *Electrochem. Soc. Trans.*, 1946.
38. Warner, A. J., Problems in the Manufacture of Ultra-high-frequency Solid-dielectric Cables, *Proc. I.R.E.*, Vol. 1 of Waves and Electrons section, p. 31W (January 1946).
39. Manufacturers' catalogs, circulars, and other data.
40. References 24, 25, 26, 28, 31, 33, 35, 39.
41. *NEMA Standards.*

Auxiliary Table. Power Factor of Insulating Materials at High Frequencies

(Approximate representative values)

Material	Frequency						
	60 cycles	1 kc	1 Mc	10 Mc	100 Mc	1000 Mc	10,000 Mc
Quartz, fused	....	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001
Polytetrafluoroethylene	0.0002	.0002	.0002	.0002	.0002	.0002	.004
Mica (muscovite)	.00024	.00024	.0002	.0002	.0002	.0002	....
Wax, mineral (best)	....	.0003	.0002	.0002	.0002	.0002	.0003
Polystyrene (best)	.0002	.0002	.0002	.0002	.0003	.0004	.0004
Polyethylene	.0005	.0005	.0003	.0003	.0003	.0004	.0004
Steatite, special	.0014	....	.0004	.00035	.00035	....	.0014
Polychlorostyrene	.0004	.0004	.0004	....	....	.0005	.0005
Glass, high silica	....	....	.0005	....	.0008	.0008	.001
Zircon porcelain	.06	.0026	.0013	.0008	.0008	....	....
Polyvinyl carbazole	....	....	.004	....	.0009	.0009	....
Glass-bonded mica	....	.003	.0014	.0011	.001	.003	....
Steatite, low loss	.0022	.0021	.002	.0015	.0015	.002	.0025
Steatite, ordinary	.015	.003	.003	.003	.0034	.0035	....
Aniline formaldehyde	.003	.004	.006	.005	.004	.004	.008
Glass, Pyrex	.006	.005	.004	.004	.005	.006	....
Rubber, hard (best)	....	.0033	.006	....	.006	.006	....
Ethyl cellulose (best)	.007	.006	.005	.006	....	.015	.025
Phenolic, mica filled (best)	.025	.014	.007	....	....	.01	....
Porcelain, wet process	.03	.02	.0086	.010	.01	....	....
Cellulose acetate	.03	.04	.03	....	....	.03	....
Vinyl resin (hard)	.015	....	.015	....	....	.007	....
Phenolic, glass base	....	....	.011	.013	.018	.02	.024
Phenolic, paper base	.05	.04	.035	.04	.05	.080	.04

ADDITIONAL INFORMATION ON SOLID DIELECTRIC MATERIALS.

Acrylic Resins and Acrylates. Designations used for thermoplastic polymers derived from acrylic acid ($CH_2 : CHCOOH$), methacrylic acid [$CH_2 : C(CH_3)COOH$], or allied materials. The most important are the methyl **methacrylates,** q.v. Acrylics may be vulcanized to reduce thermoplasticity.

Alkyd Resins. Resins derived from reaction of polybasic acids with polybasic alcohols, glycerol and phthalic anhydride being the principal materials in use. They are used chiefly for varnishes and other coatings. Trade name: *Glyptal.*

Allyl Resins. A group of plastics derived from allyl alcohol ($CH_2 : CHCH_2OH$). The resins are marketed in the form of cast transparent sheets and rods, and in the form of liquid monomers which are polymerized with the aid of a peroxide catalyst and heat, but no gases or vapors are emitted. The resin is thermosetting in nature with good electrical properties which are maintained at elevated temperatures. Index of refraction, n_D, is 1.49 to 1.51. Light transmission is over 91 per cent. Specifications: ASTM D819. Trade names: *Allymer, MR Resins, Kriston.*

Allymer * CR-39 et al. Pittsburgh Plate Glass Co. **Allyl resins.**

Alsimag.* American Lava Corp. **Steatite, cordierite,** and other ceramic bodies.

Alvar.* Shawinigan Products Corp. **Polyvinyl acetal resin.**

Amber. Yellow or orange fossil resin found on the shore of the Baltic Sea. High insulation resistance makes it an excellent insulator for electrometers.

Amphenol * 912. American Phenolic Corp. **Polystyrene** sheet, rod, and tubes.

Amphenol * 912B. American Phenolic Corp. **Methacrylate** sheet, rod, and tubes.

Aniline-formaldehyde Resins. Derived from aniline and formaldehyde and fabricated by casting and very limited molding under heat and pressure. The resin is but slightly thermoplastic, and parts are usually made by machining sheet or rod stock. Color is reddish brown. Insulation resistance is high and dielectric losses are low over a wide frequency band. Trade names: *Dilectene, Cibanite.*

Armite.* Spaulding Fibre Co. Thin vulcanized fiber (**fish** paper).

Aroclors.* Monsanto Chemical Co. **Chlorinated diphenyl resins** and oils.

Asbestos. A hydrated magnesium silicate mineral in fibrous form. Two distinct groups of minerals are described as asbestos: amphibole, or hornblende asbestos, in various subgroups, most of which contain iron and have harsh and springy fibers, relatively weak and non-flexible; and serpentine asbestos, the principal subgroup consisting of chrysotile, so-called Canadian asbestos, which has soft, fine, strong fibers, suitable for manufacture of asbestos textiles. Chrysotile asbestos is stable up to temperatures of 400 to 500 deg cent and may be useful at still higher temperatures. It is not a particularly good electrical insulator and has very high losses when moist but is valued chiefly for its fire resistance and its heat resistance. Asbestos products should not be allowed to come into contact with fine wires since corrosion may result.

* Trademark names.

Asbestos Paper. Made by felting asbestos fibers, often with some rag fibers for additional strength. It is employed normally only as a flame barrier or for heat insulation and should not be depended upon for electrical insulation.

Asbestos Textiles. May contain up to 6 per cent total iron content and as much as 2 per cent iron in a magnetic form. A higher grade known as "non-ferrous," containing less than 1.75 per cent total iron and less than 0.75 per cent magnetic iron, is available. Non-impregnated asbestos products are extremely hygroscopic. In order to avoid absorption of moisture, the products often are treated with varnishes or oils, but impregnation reduces the maximum operating temperature sharply.

Asbestos Wood. Asbestos combined with Portland cement to form dense sheets (*Transite*). Also combined with magnesia and cement in an insulating board impregnated with a black insulating compound (ebony asbestos wood).

Asbestos Ebony.* Johns-Manville Co. Asbestos fiber and binding cement impregnated with compound.

Asphalts, Natural. Native asphalts are roughly divided into relatively pure deposits containing less than 10 per cent mineral matter and those containing a large amount of mineral matter. Both types of deposit contain water, but in the latter group the water is often in emulsified form. The water content may be as high as 40 per cent. It is very difficult to separate any but the largest particles of mineral matter by heating. Asphalt is used for potting and impregnating compounds, in the manufacture of varnishes and japans, and for the insulating covering of cables. It is closely related to petroleum asphalt and asphaltites, which are used for the same purposes. Refined Trinidad asphalt melts at 87 deg cent and contains about 38 per cent mineral matter. Bermudez asphalt melts from 57 to 87 deg cent and contains about 4 per cent mineral matter.

Asphalts, Petroleum. A "rubberlike" asphalt of almost any desired melting point up to 150 deg cent. The residue from the distillation of asphaltic or mixed-base petroleums. Sometimes called residual pitches or asphalts. Have greater purity and uniformity than natural asphalts. Used for potting, impregnating, etc. Some residual asphalts weather very badly when exposed to sunlight. Petroleum asphalt may be "blown" with air and steam to oxidize the asphalt partially and to increase the melting point. They also may be modified by adding rosin derivatives to increase the fluidity at high temperatures without drop in melting point.

Asphalts, Sulfurized. Sulfur has a hardening action on asphalt, similar to vulcanization of rubber in some respects. Used in corona-resisting wire insulation.

Asphaltites. Asphaltlike substances but much harder; have melting points above 120 deg cent. Not as soluble in petroleum hydrocarbons. The most important varieties are: **gilsonite,** m.p. 130 deg cent; glance **pitch** or manjak, m.p. 160 deg. cent; and grahamite, m.p. 175 deg cent Another group of asphaltites containing so-called pyrobitumens are practically infusible. Elaterite, wurtzilite, albertite, and impsonite are members of this group and are almost insoluble in the usual solvents. The first group is used in the manufacture of varnishes and japans.

Bakelite.* Bakelite Corp. **Phenolic, cellulose acetate, polystyrene, urea,** and other resins, plastics, and molding powders.

Balata. A rubberlike natural material similar to gutta-percha and used for similar purposes. M.p. 150 dec cent. Dielectric constant 2.6 to 3.5, depending on purity. May be deresinified to improve properties for use in submarine cables. See Kemp, *J. Franklin Inst.*, Vol. 211, No. 1, p. 37.

Beeswax. A white to yellow insect wax. Plastic at 30 deg cent. Melts from 62 to 64 deg cent. A good electrical insulator. Has a large negative coefficient of volume resistivity; is bleached by sunlight and turns brown with age.

Beetle.* American Cyanamid Co. **Urea formaldehyde molding powders.**

Buna N. Synthetic rubber made from butadiene and acrylonitrile by emulsion polymerization. Heat-, oil-, and solvent-resistant, but electrical properties are not good. Trade names: *Hycar OR*, *Perbunan*, *Chemigum*.

Buna S. Common type of synthetic rubber made from butadiene and styrene by emulsion polymerization. Rubber made in government plants is known as GR-S. Insulation fully equal to that of natural rubber in electrical properties can be made from Buna S although the mechanical properties are not quite so good. Its general behavior is quite similar to that of natural rubber.

Butacite.* E. I. du Pont de Nemours & Co. **Polyvinyl butyral plastics.**

Butvar.* Shawinigan Products Corp. **Polyvinyl butyral plastics.**

Butyl. A synthetic rubber made from isobutylene and butadiene by polymerization at low temperatures with a catalyst. It has excellent electrical properties, good resistance to heat, ozone, and corona, and low permeability to gases but is inferior to natural rubber in strength. When made in government plants it was known as GR-I.

Casein. Prepared from skim milk by rennet treatment and hardened by soaking in formaldehyde. Before hardening it can be extruded and pressed; after hardening it is readily machined. It softens in hot water at 100 deg cent and can be blanked or molded to a limited extent. It is not a particularly good electrical insulator but is occasionally useful for small parts. It is readily colored either before or after fabrication.

Catalin.* Catalin Corp. **Cast phenolic resins.**

Cellophane.* E. I. du Pont de Nemours & Co. Regenerated **cellulose film,** lacquered to reduce moisture transmission.

Celluloid.* Celanese Plastics Corp. **Cellulose nitrate.**

Cellulose Acetate. A thermoplastic prepared by treatment of cotton linters or other cellulose with acetic anhydride and acid and the addition of suitable plasticizers. Variation of processing and plasticizer gives a wide range of molding flows and of the heat resistance, flow mobility, and other properties of the product. Available in all colors and in the forms of film, sheet, rod, and molding powders for

---

* Trademark names.

compression and injection molding or for extrusion. Films and sheet of electrical grade are non-corrosive to fine copper wires even under conditions favorable to electrolysis and are superior for coil construction. Some molded materials contain volatile plasticizers and shrink in time or with heating. The heat distortion point of many grades is also quite low. Cellulose acetate is not attacked by oils but is dissolved by ketones, esters, and other solvents. Specifications for sheets: ASTM D786; for molding powders, ASTM D706. Trade names: *Fibestos, Lumarith, Plastacele, Tenite I.*

Cellulose Acetate-butyrate. Similar to cellulose acetate except that it is a mixed ester of cellulose. Less plasticizer is required to obtain molding flow, and heat resistance is improved slightly. It is widely used in tape form for wire insulation and as a molding powder. Specifications for molding powders: ASTM D707. Trade name: *Tenite II.*

Cellulose Nitrate. Also known as pyroxylin and *Celluloid.* It is made by nitrating cotton linters or wood pulp and adding suitable plasticizers such as camphor. Basically highly inflammable, but appropriate plasticizers reduce the inflammability. Very tough and is water- and chemical-resistant. Marketed as sheets, rods, tubes, and molding compositions. Specifications for sheet, rod, and tube: ASTM D701. Trade names: *Celluloid, Pyralin, Nitron, Nixonoid.*

Cellulose Propionate. A new plastic similar to cellulose acetate but with improved impact strength and excellent dimensional stability. Good electrical properties and low water absorption. Trade name: *Forticel.*

Celeron.* Continental-Diamond Fibre Co. Macerated-fabric-base **phenolic moldings.**

Ceramics. Electrical ceramics include materials sintered, fused, or fired at high temperatures, such as **porcelain, steatite, cordierite, glass, glass-bonded mica,** titanium and zircon ceramics, q.v. Ceramics are characterized by heat resistance, permanency of dimension, low water absorption, low thermal expansion, and excellent electrical properties. Care is required in mounting ceramic parts since the modulus of elasticity is high and a slight bending develops high stresses. With the exception of glass-bonded mica, ceramics are not machinable by normal methods and are normally cast, molded, pressed, or extruded and machined in the unfired state and then fired at a high temperature. Since firing shrinkage is around 12 per cent, tolerances are not close, usually of the order of 1 per cent of the dimension. A limited amount of grinding can be done after firing.

Ceresin. A white or yellow wax, with exceptional dielectric properties. Used extensively alone and mixed with other waxes. M.p. 65 to 67 deg cent. Soluble in oils and petroleum distillates. Very water-resistant and has high surface resistivity.

Cerex.* Monsanto Chemicals Co. **Polystyrene copolymer.**

Chatterton's Compound. Fusible composition of gutta-percha, rosin, and tar used in submarine cable construction for sealing cable ends, etc.

Copaline.* American Phenolic Corp. Solid-dielectric r-f cables.

Cordierite Ceramics. Consist principally of the crystal cordierite, a magnesium aluminum silicate, and are characterized by a very low coefficient of thermal expansion and a power factor much higher than that of steatites. Trade name: *Alsimag 72* and *202.*

Co-ro-lite.* Colombian Rope Co. Sisal-fiber-reinforced **phenolic materials.**

Corprene.* Armstrong Cork Products Co. Cork-loaded **neoprene** sheet in various grades.

Dilectene.* Continental-Diamond Fibre Co. **Aniline formaldehyde** sheets and rods.

Dilecto.* Continental-Diamond Fibre Co. **Laminated phenolic** sheet, rod, and tubes.

Dri-film.* General Electric Co. Silicone treatment for **ceramics.**

Durez.* Durez Plastics and Chemicals, Inc. **Phenolic resins** and **molding powders.**

Durite.* Durite Plastics, Inc. **Phenolic** and **furfural resins.**

Ebonite. Another name for **hard rubber.**

Empire.* Mica Insulator Co. **Varnished cloth,** sheet, and tape. Silk, rayon, *Fiberglas,* 2 to 40 mils thick.

Enamel. General term for a substance producing a colored glossy coating. Organic enamels are made by pigmenting or coloring lacquers or varnishes, and inorganic enamels by pigmenting low-melting glasses. (See discussions below.)

Enamel, Varnish. Pigmented varnishes or varnishes colored by asphalts or colored resins. Varnish enamels may either dry in air in 4 to 16 hr, depending on the type, or be compounded to dry by baking from 2 to 8 hr at 100 to 150 deg cent. The baking types usually have better adhesion and hardness and higher dielectric properties. The japans are varnish enamels made with asphalts for baking.

Enamel, Vitreous. A silicate or borosilicate glass with the melting point lowered by the addition of various fluxes, such as metallic oxides and salts. Some metallic oxides are used for coloring the enamel and others to increase the opacity. The enamel is usually applied in two or more coats when freedom from pinholes is desired. The same composition may be used for all coats in some cases, although specially formulated ground coats are usually superior. The enamel coats are applied by dipping or spraying, using a suspension of the finely ground enamel, known as frits, in a clay-and-water medium. For flat work the powdered frits are sometimes sieved onto the work in the so-called dry process. The coats are "fired" at temperatures up to 850 deg cent for periods of 3 to 30 min. The temperature and time are critical for a given composition of enamel and for the size of the enameled object. Some enamels have very satisfactory dielectric properties and may even be used for the dielectric of small adjustable capacitors, but the major electrical use is the manufacture of vitreous enamel resistors. For this use the formula is carefully adjusted to obtain a coefficient of thermal expansion suitable for the wire and coil forms, to prevent fracture of the wire with cycles of heating and cooling.

Enamel Wire. Enamel for insulating wire is usually a varnish enamel. A hard, tough, and flexible covering is produced with a high dielectric strength. The enamel thickness varies from 0.0001 to 0.0015 in., depending on the wire size. Enamel is suitable for continuous operation at 105 deg cent. Wire enamels vary slightly in resistances to oils and solvents. The better grades are suitable for con-

* Trademark names.

tinuous use in mineral oil up to 80 deg cent; 48 hr at 105 deg cent should have no visible effect. Most varnish and lacquer solvents will attack wire enamel to some degree if the contact is prolonged. It is important, therefore, to expel the solvents from varnish impregnation or cementing processes with reasonable promptness. It is necessary in most cementing processes for the solvents to attack the enamel slightly in order to secure a good bond. Wax impregnation has practically no effect on the enamel, but impregnation with asphaltic compositions may cause a severe attack on the enamel. The enamel is moderately resistant to abrasion and pressure, and wire may be random wound for some uses where a few shorted turns are allowable. It is also used successfully on core plates where it is subjected to considerable pressure and is occasionally filled with mica dust to improve the resistance to pressure.

Ethocel.* Dow Chemical Co. Ethyl cellulose molding powder.

Ethofoil.* Dow Chemical Co. Ethyl cellulose film.

Ethyl Cellulose. Prepared by replacing the hydrogen of cellulose hydroxyl groups with the ethyl group by means of ethyl chloride or ethyl sulfate. It is thermoplastic and is tough and flexible through wide ranges of temperature. It is very stable to heat and light and has excellent electrical properties. It is available in "hot-melt" compounds, in casting compositions (*Thermocast**), in films, as non-rigid sheets and extrusions, and as molding compositions. The electrical properties are well retained under moist conditions and actually may improve with increasing temperatures, contrary to the general rule. Specifications for non-rigid: ASTM D743; for molding: ASTM D787. Trade names: *Ethocel, Lumarith EC, Chemaco, Hercules EC.*

Fiber, Vulcanized. Manufactured by treating rag paper with zinc chloride, pressing into sheets, and thorough washing. Fiber will absorb water up to about 60 per cent if immersed for a sufficient length of time and will increase to nearly double the thickness. Solvents and oils have practically no effect on fiber. Fiber has a natural moisture content of 5 to 6 per cent at 40 to 60 per cent relative humidity, which will decrease at low humidities and increase at high humidities. Heating at 80 to 100 deg cent for long periods will dry out and warp the fiber and impair the flexibility and toughness. From 100 to 170 deg cent the drying is very rapid, and the fiber will become brittle in a few hours. Above 170 deg cent the material will char on long heating. It chars in a short time at 200 deg cent. The maximum safe operating limit is about 150 deg cent. Fiber is made in gray, black, red, and white, but there is little difference in the properties due to the colors of the same grade of stock. Grade differences are due to selection of rags, increased pressure in order to increase density and hardness, and more careful processing. The highest grade is the electrical grade made in thin sheets for use in slot insulation, etc., and commonly known as "fish paper." This grade has exceptionally high dielectric and mechanical strength and a density of about 1.3. Hard fiber (bone fiber, horn fiber) has a density of about 1.3, and is considerably harder than the commercial grade. The commercial grade has slightly lower mechanical strength but about the same dielectric strength and shows considerably higher water absorption on 1-hr immersion. A grade especially made for forming and swaging contains glycerin to soften the material by retaining moisture. It can be swaged, spun, and formed readily for bushings, grommets, etc. All grades are readily machinable and can be punched and also formed dry to a limited extent. For more severe operations, the sheets can be soaked in water and formed in heated dies. The fiber bends best parallel to the grain.

Fiberglas.* Owens-Corning Fiberglas Corp. Glass yarn and textiles.

Fibestos.* Monsanto Chemical Co. Cellulose acetate.

Fibron.* Irvington Varnish and Insulator Co. Flexible plastic products. Polyethylene tape and tubing.

Fish Paper. Common name for superior grade of thin vulcanized fiber, also known as tarpon paper, leather paper, leatheroid, and fiberoid. It is usually supplied in a dark gray color.

Flamenol.* General Electric Co. Polyvinyl chloride wire insulation.

Formex.* General Electric Co. Polyvinyl formal magnet wire insulation.

Formica.* Formica Insulation Co. Phenolic laminates.

Formvar.* Shawinigan Products Corp. Polyvinyl formal.

Forticel.* Celanese Plastics Corp. Cellulose propionate plastic.

Fortisan.* Celanese Corp. of America. High-strength regenerated cellulose yarns.

Fullerboard. Another name for pressboard.

Furfural Resins. Phenol and furfural react to form resins similar to phenol-formaldehyde resins but generally dark in color. They are used for transfer molding and cold molding. They withstand high temperatures and have comparatively good arc resistance. Trade name: *Durite.*

Gelva.* Shawinigan Products Corp. Polyvinyl acetate.

Geon.* B. F. Goodrich Chemical Co. Polyvinyl chloride and vinyl vinylidene chloride.

Gilsonite. Variety of natural asphalt. M.p. 122 to 188 deg cent.

Glass. Physically, glass is an amorphous, undercooled liquid composed of silica and metallic silicates and hence has no crystal structure or sharp melting point. The plastic nature of glass at elevated temperatures permits fabrication by drawing, blowing, and pressing, as well as by casting. Only simple shapes can be molded. All glass must be carefully annealed to prevent residual stresses which reduce the strength, or surface-chilled in a controlled manner, as in so-called tempered glass, so that the residual stresses will increase the strength. Since glass has a high elastic modulus and no internal structure to interrupt stress patterns, it is extremely sensitive to stress concentrations. Glass parts must be carefully designed, and surface damage must be avoided. Exclusive of glasses designed for sealing purposes, lime glass, lead glass, borosilicate glass (*Pyrex* types), and high-silica glass (Vycor) are the principal electrical glasses. The borosilicate and high-silica types have lower coefficients of thermal expansion, giving improved resistance to thermal shock. Electrical glasses have high dielectric strength and volume resistivity and low power factor. These properties depreciate with rising temperature, and at temperatures from 150 to 200 deg cent a rapid rise in power factor and decrease in dielectric strength

* Trademark names.

begin. The high-silica glasses are superior in this respect. The surface of glass is readily wetted by water so that the surface resistivity is seriously reduced by relative humidities above 70 per cent. This is perhaps the most serious defect of glass as an electrical insulator, but recent work has indicated that leakage may be considerably reduced by treatment with silicones, q.v. The *Multiform* process, in which glass is powdered, pressed to shape, and fired like porcelain, makes possible the production of complicated parts to tolerances of 1 to 2 per cent and permits the use of high-silica glasses with very low losses and low thermal expansion. Recently developed techniques permit the firing of metallic coatings on glass, which may be used as circuit conductors, for soldered connections, or hermetic sealing by soldering. By alteration of the composition of glass, the thermal expansion may be matched to certain metals so that dependable seals resistant to thermal shock can be produced. Lead-through seals are available in single or multiple form, which may be soldered in a metal container to give a hermetic seal.

Glass Textiles. Glass drawn into thin filaments (2 to 3 $\times$ 10^{-4} in. diameter) has an enormous tensile strength (4 to 5 $\times$ 10^5 lb per sq in.), owing to an absence of shearing stresses. Glass textiles (*Fiberglas* *) are made from glass yarn of two basic types: continuous-filament yarns made by hot drawing of filaments from glass "marbles" in a special machine, and staple yarns made from staple fiber produced by steam drawing of molten glass into fibers varying from 4 to 18 in. in length. The individual fibers are lubricated and combined into yarns of various constructions. Glass textiles are available as tapes, cloth, sleeving, cords, and yarns for serving or braiding. Outstanding uses for glass textiles have been: (1) high-strength plastic laminates; (2) fireproof braiding for insulated wire; (3) varnished glass cloth; (4) serving for magnet wire. Advantages of glass textiles in these and other services are greater resistance to heat, longer life, non-inflammability, increased moisture resistance, and high mechanical strength.

Glass-bonded Mica. Ground mica bonded with a low-melting glass, chiefly lead borate, and sometimes with the addition of cryolite (sodium aluminum fluoride). The material is hot-pressed at 600 to 700 deg cent to the required form. Certain types can be injection-molded. Metal inserts can be molded in place, in some cases for hermetic sealing purposes. It is also possible to cast aluminum around the material or to use it as an insert in plastic molded parts. It can be readily machined with carbide tools, or with ordinary tools for small quantities, to close tolerances. Glass-bonded mica has excellent electrical properties, good mechanical properties, low coefficient of expansion, and stability in dimensions up to 300 deg cent. Electrical properties do not deteriorate rapidly with rising temperature or under moist conditions, except that the surface resistivity of grades containing fluorides may be very low at high humidities. Under condensation conditions, the fluoride constituents are dissolved and may corrode metal parts. Polishing and waxing the surfaces will eliminate this condition but will reduce the arc resistance. Special varieties of glass-bonded mica (*Mycalex K* *) are available with controlled dielectric constants between 8 and 20. Specifications: Army-Navy Specification JAN-I-10, Grade L-3 or L-4. Trade names: *Mycalex*, *G. E. Mycalex*, *Mycroy*, *Turx*.

Glyptal.* General Electric Co. Alkyd resins and products such as varnishes, cements, varnished cloth, etc., made therefrom.

Gummon.* Garfield Mfg. Co. Asbestos coal-tar moldings.

Gutta-percha. A grayish-white to brown plastic substance but not elastic like rubber. Can be molded under pressure at 60 to 100 deg cent and melts from 120 to 140 deg cent. It vaporizes above 190 deg cent. Partly soluble in ether, carbon tetrachloride, benzol, chloroform, and carbon bisulfide; insoluble in water. Can be vulcanized with sulfur or sulfur chloride like rubber, forming a hard substance, but it is nearly always used unvulcanized. It is rather easily oxidized in the air and becomes brittle and yellowish gray. It is principally used for the insulation of submarine cables and is generally applied uncompounded by a tubing machine or in strips like rubber. The power factor of gutta-percha is maximum at room temperatures, so that the dielectric loss in actual service at sea-bottom temperatures is quite low. If the insulation is prepared with about 1.5 per cent moisture content, which is close to the saturation value in sea water, the constants do not change much in service. The life is very satisfactory under water but is not very satisfactory in air.

Halowax.* Bakelite Corp. Chlorinated naphthalene liquids and waxes for impregnating.

Hard Rubber. See **Rubber, hard.**

Hemit.* Garfield Mfg. Co. Cold-molded refractory materials.

Herculite.* Pittsburgh Plate Glass Co. Tempered glass products.

Hycar.* Hycar Chemical Co. Synthetic rubber in various grades, distinguished by suffix letters and numbers.

Insurok.* Richardson Co. Phenolic and urea laminates.

Isolantite.* Isolantite, Inc. Steatite ceramics.

Jute. A long bast fiber employed in cordage and rough textiles. Considerably used as a filler and core in cords and cables. Commercial jute often is softened and rendered less brittle by impregnation with mineral oil. Jute loses its strength when damp.

Kaolin. Also called china clay. An aluminum silicate clay free from iron, used in the manufacture of white porcelain. Valuable as a packing material around heating coils, etc.

Kriston.* B. F. Goodrich Chemical Co. Allyl monomer.

Lamicoid.* Mica Insulator Co. Phenolic laminates.

Laminates. Layers of paper, cloth, or glass cloth impregnated with resin and pressed under heat and high pressures. Resins are usually thermosetting, but thermoplastic resins have been used. Low-pressure laminates use resins which give off little or no gas or vapor during curing, and pressure just sufficient to hold the mass in contact is required. Very large, shaped parts can be produced by employing inflated or evacuated rubber bags to supply the low pressures needed. Knitted cloths are often used to allow stretching where required. Contact laminates can be made with still lower pressures (as

* Trademark names.

low as 1 lb per sq in.). Trade names of materials employed for contact laminating are: *Laminac,** American Cyanamid Co.; *Thalids,** Monsanto Chemical Co.; *Bakelite Copolymer Resins,** Bakelite Corp.; *Selectron,** Pittsburgh Plate Glass Co.; *Vibrins,** Naugatuck Chemical Co. The electrical properties of many of these low-pressure laminates are excellent. They were employed for such applications as radomes during the war. Properties of sample laminates: dielectric constant 3.54 to 4.59 at 1 Mc; power factor 0.0075 to 0.0105 at 1 Mc.

Latex. An emulsion of rubber, synthetic rubber, or synthetic resin in water, depositing a solid film on evaporation.

Lava.* American Lava Corp. Mineral talc machined to shape and fired at high temperatures.

Lavite.* D. M. Stewart Mfg. Co. **Steatite ceramic.**

Lenoxite.* Lenoxite Div., Lenox, Inc. **Steatite ceramic.**

Lignin. Lignin is the binding material in wood. Two types of plastic are made from lignin. In one type the whole wood is used; steamed chips are exploded by sudden pressure release and are pressed into boards under high pressure (*Masonite,** *Benalite,** Masonite Corp.). In the other type the separated lignin, usually a byproduct of paper manufacture, is combined with other materials, such as amines, furfural, or phenol, to form a thermosetting resin which may be combined with various fillers, or is used to impregnate paper which is hot-pressed into laminated boards (*Lignolite**). Lignin also is used as an extender for phenol-formaldehyde molding compounds.

Lignolite.* Marathon Chemical Co. **Lignin** plastic sheets.

Loalin.* Catalin Corp. **Polystyrene.**

Lucite.* E. I. du Pont de Nemours & Co. **Methyl methacrylate** sheet and molding powders, also in heat-resistant grades.

Lumarith.* Celanese Celluloid Corp. **Cellulose acetate** and **ethyl cellulose** products.

Lustron.* Monsanto Chemical Co. **Polystyrene** molding powders.

Magnesium Oxide. Compressed magnesium oxide is used in the insulation of heating units and, in Europe, for heat-resistant coaxial conductors (*Pyrotenax**). Single or multiple conductor cables are made by packing magnesium oxide preforms and the conductors inside a copper tube and drawing the assembly to a smaller size. As a r-f coaxial line the losses are higher than those of polyethylene insulation, and the ends must be well sealed against moisture. It is electrically smooth compared to an insulator-spaced air line.

Makalot.* Plastics Div., Interlake Chemical Co. **Phenolic resins** and compounds.

Masonite* Die Stock. Masonite Corp. Exploded wood fiber, densified under high pressure.

Melamine-formaldehyde. Thermosetting resins prepared by reaction of formaldehyde, melamine, and sometimes dicyandiamide; the latter two are derived from calcium cyanamid. These resins are heat- and arc-resistant and have excellent electrical properties and low water absorption. Alpha cellulose, chopped rag, and mineral fillers are used in various compounds. Specifications: ASTM D704. Resin trade names: *Melmac, Resimene, Plaskon Melamine.*

Melamine Glass Laminates. These laminates are characterized by high arc resistance and great mechanical strength. They are also heat-resistant and burn with some difficulty. The fumes from the burning laminate are said to be less toxic than those from phenolics; hence these laminates were used for combat-vessel equipment. The high-frequency properties are not outstanding and are not controlled in production. Specifications: Joint Army-Navy Spec. JAN-P-13 Type GMG.

Melmac. American Cyanamid Co. **Melamine-formaldehyde resins** and molding powders.

Methacrylates. These resins are members of the acrylic or acrylate resin group. The most important member is methyl methacrylate. This plastic is produced by the reaction of acetone and hydrogen cyanide to form acetone cyanhydrin, which is allowed to react further with methyl alcohol to produce methyl methacrylate monomer [$CH_2 : C(CH_3)COOCH_3$]. This monomer is polymerized by the aid of peroxide catalysts and heat. The polymers are characterized by great optical clarity, light transmission of 92 per cent, high refractive index of 1.48 to 1.51, stability to light and weather, and good mechanical and electrical properties. Arc resistance is high; vapor from the plastic actually tends to quench arcs. Power factor and dielectric constant decrease with increasing temperature and frequency instead of exhibiting the normal increase. Methyl methacrylate in common with other thermoplastics has a low heat distortion point. Heat-resistant grades are available that will withstand boiling in water. Specifications for sheet, rods, and tubes: ASTM D702; for molding compounds: ASTM D788. Trade names: *Plexiglas, Lucite.*

Mica. A group of natural complex aluminum silicates with highly developed basal cleavage into thin, tough, flexible laminae. It is probable that if sufficient care were taken it could be split into thickness approaching molecular dimensions. Owing to the fact that blocks are very expensive, mica is usually split and punched into parts for capacitors and spacers. For other uses the flakes are cemented together with adhesives to make "built-up" or "pasted" mica, which forms flexible or rigid sheets, depending on the binder. Flake or dust mica is combined with resins or glasses to form simple molded shapes as well as rods and sheets. There are several varieties of mica, including:

Biotite—iron mica, black mica	Muscovite—potassium mica, common mica
Paragonite—sodium mica	Phlogopite—magnesium mica, rhombic mica
Lepidolite—lithium mica	
Lepidomelane—iron mica	

but only the last two are used for electrical work. Muscovite comes in three colors: white and ruby, both of which are superior grades, and green, which is inferior electrically and mechanically to clear grades of white and ruby. Muscovite mica is in general superior electrically and mechanically to phlogopite, but phlogopite has superior heat resistance. Phlogopite ranges from a deep amber color to dark amber and milky white. It is not so readily split as muscovite; it is softer, and is lower in mechani-

* Trademark names.

cal strength. Phlogopite mica does not lose water up to temperatures of 800 to 900 deg cent, and some grades will resist 1200 deg cent without complete disintegration. For this reason it is valuable for spacing the elements in vacuum tubes and in heating devices. The maximum operating temperature is best limited to 1000 deg cent, and that of muscovite to 500 deg cent. The power factor and resistivity of phlogopite are much worse than those of muscovite, although the dielectric strength is nearly the same. Stained muscovite and all phlogopite are unsuitable for use in capacitors where low power factor is required. Mica, in general, does not decrease in dielectric strength with frequency as fast as most dielectrics. This fact, together with its low power loss, enables carefully designed and built mica capacitors to operate at extremely high frequencies. The defects occurring in mica are air bubbles, stains, and spots. Mica is graded according to freedom from defects as follows: highest-grade mica is clear and free from all defects; second highest grade has air bubbles between laminae; stained mica sometimes has some iron stains present; spotted mica is badly stained and usually has inclusions of other minerals. Mica splittings are graded for size according to the largest usable rectangular area. For grading methods, see ASTM D351. The dielectric strength of mica is considerably reduced by air or moisture between the laminae, but the flexibility is somewhat increased. Specifications for block mica and films: ASTM D748; for electrical tests: ASTM D351.

Mica, Reconstructed or Pasted. Flake mica is bound and pressed together with shellac, gum, asphalt, or synthetic resin varnishes and milled to thickness to form sheets which may be punched and sheared. Some grades are flexible and may be formed to a limited extent cold. Others use a thermoplastic binder and can be formed to quite intricate shapes at 100 deg cent. Hard grades contain as little as 3 per cent binder and do not compress appreciably. See also **Mica cloth** and **Mica paper.** Specifications for materials: NEMA Standards 39-55; for testing methods: ASTM D352.

Mica Cloth. A composite insulation of high dielectric strength used for insulating transformers and field windings.

Mica Paper. Flake mica cemented between sheets of glassine, rice, kraft, or express paper. Mica also is combined with asbestos paper or fibers to form composite insulations.

Mica Plate. Another name for **reconstructed mica** sheets.

Micabond.* Continental-Diamond Fibre Co. **Reconstructed mica** tape, tubes, and sheets.

Micanite.* Mica Insulator Co. **Reconstructed mica** products.

Micarta.* Westinghouse Electric Corp. **Phenolic laminates.**

Minerallac.* Minerallac Electric Co. Fusible **asphalt** compounds.

Molded Compounds. *Hot-molded* products are formed in molds or platens heated to a temperature sufficient to cause the binder to flow, cementing the particles and producing a pure smooth binder surface which lowers water absorption and increases surface resistivity. Compounds in which binder hardens under heat are called thermosetting; those in which binder becomes plastic are called thermoplastic, and molds must be cooled before the article is removed. Some thermosetting binders are synthetic resins of the phenol **formaldehyde, urea formaldehyde,** or **melamine** type. Some thermoplastic binders are **shellac, cellulose nitrate** and acetate, **vinyl resins,** mixtures of **asphalts** and hardened **rosin,** copals, **casein resins,** sulfur chloride phenol resins, cumarin resins, **polystyrene,** and **methacrylate.** Fillers may be either fibrous or powdered, wood flour being the commonest. Cotton, silk flock, or threads are used to improve resistance to impact, and macerated cloth to give high impact resistance, with thermosetting compounds. Asbestos is used for heat-resistant products, mica to obtain low power factor, and ground flint, china clay, silex, stone, etc., to cheapen the article. Compounds can be hot-molded in great varieties of shapes with thin sections, metallic inserts, threads, etc. Tolerances can be held to within plus or minus 0.003 in. per in. when not depending on mold closure. *Cold-molded* products are formed under pressure and subsequently baked for periods of from a few hours to a week at temperatures of 150 to 300 deg cent. There is little flow of the binder, and the surface is not very smooth and depends on the fineness of the filler. Pieces distort slightly in baking, and the accuracy of dimensions is much lower than in the hot-molded process: it amounts to plus or minus 0.009 to 0.015. Binders are thick **varnishes, asphalts,** tung and linseed oils, anthracene oils, and various gums or varnish resins in suitable solvents. Cold-molded compounds are subdivided into refractory and non-refractory, according to the degree of heat resistance. In general, close tolerance, loose mold pieces, and threaded parts increase molding cost. For low cost the parts should be designed with adequate radii, and with no projections, indentations, or holes which require loose pieces in the molds.

Multiform Glass.* Corning Glass Works. Powdered **glass,** pressed and fired.

Muscovite. Variety of mica suitable for electrical insulation. See **Mica.**

Mycalex.* Mycalex Corp. **Glass-bonded mica,** sheet, rod, and molded.

Mycalex,* G. E. General Electric Co. **Glass-bonded mica,** sheet, rod, and molded.

Mycroy.* Electronic Mechanics, Inc. **Glass-bonded mica.**

Neoprene. Produced by emulsion polymerization of chloroprene, which is derived from acetylene. Polychloroprene is the chemical name for neoprene but neoprene generally is used even though it is the trade name for the E. I. du Pont de Nemours & Co. product. Neoprene can be vulcanized like rubber but sulfur is not always required. There are many types of neoprene distinguished by suffix letters; some of the types are copolymers with nitriles. Neoprene made in government plants was called GR-M; freeze-resistant neoprene, FR; general purpose, GN; low oil swelling and low gas diffusion, ILS, etc. All types are outstanding in oil, ozone, and sunlight resistance and will not support combustion. Neoprene may be compounded to give power factors of 1 per cent and resistivities of 10^{12} ohm-cm, but many compounds are poor electrically. Neoprene may also be formulated for high heat resistance or for good abrasion resistance. One application of great value is the jacket of portable cords and cables. Neoprene also may be made with low specific resistivity for use as electrostatic shields or for potential control. It also is an excellent gasket material.

Nitrocellulose. See **Cellulose nitrate.**

---

* Trademark names.

Nitron.* Monsanto Chemical Co. **Cellulose nitrate.**

Nixonite.* Nixon Nitration Works. **Cellulose acetate.**

Nixonoid.* Nixon Nitration Works. **Cellulose nitrate.**

Nylon.* E. I. du Pont de Nemours & Co. **Polyamide** products of all kinds.

Ozokerite. A natural mineral wax, amorphous and black to dark brown. When bleached, it is white to yellow or brown. Melting point is 70 to 80 deg cent. The purified wax is known as **ceresin.**

Panelyte.* St. Regis Paper Co. **Phenolic laminates.**

Paper. Paper for insulating purposes should be as free as possible from all chemicals and from conducting particles and should be strong mechanically. Paper changes its moisture content very rapidly with changes of atmospheric humidity. Single sheets of thin paper will come to equilibrium in as little as 15 minutes, so that all testing, both mechanical and electrical, is done best under conditions of controlled humidity and temperature. The principal types of paper of interest are those following. Rag papers made with a minimum of chemicals and short "cooks" give strong paper with heat resistance somewhat improved over that of chemical wood papers. Manila papers, made from manila fiber or from old rope, etc., sometimes with cotton or linen rags, have been the standard cable papers for some years because of high mechanical and electrical properties. Kraft papers can now be made, however, with equal or superior properties; kraft papers properly made are very strong, have excellent dielectric properties, and are cheaper than rag or manila paper. Glassine or onion-skin paper is a highly beaten sulfite stock with fair mechanical and electrical properties. It does not take impregnation well, however. High-density, well-beaten, heavily calendered stocks in the thicker sizes are much used as insulating strips. See **Pressboard.** Material intermediate between paper and boards in density or thickness is known by various names—express paper (chemical wood fiber), rope paper (from old ropes), etc.

Paragutta. Submarine cable insulation compounded of purified gutta hydrocarbon and deresinified rubber.

Perbunan.* Standard Oil Co. of N. J. Synthetic rubber: a copolymer of butadiene and acrylonitrile.

Petrolatum. Comes in liquid, soft, and hard grades. The soft form is similar to vaseline and is used extensively as a paper-cable-impregnating material. M.p. 50–55 deg cent. The electrical properties vary with purity.

Phenol Fiber. General term for paper-base **phenolic laminates.**

Phenolic Insulating Materials. Obtainable in two principal forms: molded parts; and laminated sheets, rods, or tubes with paper or fabric base. For molding compounds, the resins are combined with the desired fillers, either by working on rolls or by coating the filler particles with a varnish and drying (see discussion of **molded compounds**). For laminated products, sheets of paper or fabric are coated or impregnated with varnish and pressed hydraulically between heated platens. Rods are wound on small mandrels which are removed and the roll is cured in a mold. This leaves a weak center section, and for some purposes rod turned from sheet stock is preferred, although it does not machine as well and splits more readily. For molded tubes the impregnated paper is wound on mandrels and the assembly cured in heated molds. Since the pressure, of course, is not radial unless an expanding mandrel is used, the seams in molded tubes are weak and tend to split apart. To overcome this, a rolled tubing is manufactured which is cured by heated rolls during the winding. Since the pressure is limited in this process, the electrical properties generally are not equal to molded tubing but the mechanical properties are superior. Laminated phenolic insulation is hard, tough, and rigid, but more elastic than equivalent molded compounds. It is infusible and resists temperatures up to 125 deg cent, but becomes slightly more brittle upon cooling after continuous operations above 90 deg cent, and usually shrinks somewhat more than it had expanded. A slight softening is noted while the compound is hot, of which advantage is taken to reduce breakage in punching operations. Stress applied while hot causes a slight permanent set, and a limited forming is thus possible. The dielectric properties are not so good as those of hard rubber, but, mechanically, phenolic insulating compounds are superior and do not corrode metals or deteriorate with age. Specifications for laminates: NEMA Standards, and ASTM D709; for molding compounds ASTM D700.

Phenolic Resins. Phenol and various other phenolic substances, such as cresol, will condense and polymerize with aldehydes under the influence of heat and a suitable catalyst. Formaldehyde and hexamethylene tetramine are the commonest substances employed to react with phenols or cresols. The reaction is catalyzed by ammonia, alkalies, acids, and other agents. The reaction proceeds in two or more stages. In the first stage the resin is fusible and soluble in acetone and other solvents. Upon further heating the resin becomes infusible and practically insoluble. This second stage is the base of thermosetting molding compounds and of some phenolic laminating or baking varnishes.

Phenolite.* National Vulcanized Fibre Co. **Phenolic laminates.**

Phlogopite. Variety of **mica,** q.v.

Piccolastic.* Pennsylvania Industrial Chemical Corp. Substituted **styrene** polymers.

Plaskon.* Plaskon Div., Libby-Owens-Ford Glass Co. **Urea** or **melamine** molding compounds.

Plastacele.* E. I. du Pont de Nemours & Co. **Cellulose acetate** products.

Plax.* Plax Corp. **Polystyrene.**

Plexiglas.* Rohm and Haas Co. Methyl **methacrylate** products.

Pliolite.* Goodyear Tire and Rubber Co. **Cyclicized rubber** thermoplastic **resin.**

Polectron.* General Aniline & Film Corp. **Polyvinyl carbazole** resin.

Polyamides. Thermoplastic resins formed from dibasic acids and diamines. *Nylon,* the most important, is formed from adipic acid and hexamethylene diamine. It is characterized by extraordinary strength and toughness, and by a high degree of resistance to solvents and chemicals. The electrical characteristics of nylon are good but not so outstanding as its mechanical properties. It has been used

* Trademark names.

successfully for thin-wall coil forms and for thin-wall jacketing of assault wire. Its strength and fungus-resisting properties have led to its use for many military applications in connection with parachutes, aircraft, guy, mooring, and tow ropes, cords, etc. It is available as yarn, monofilament, and molding compound. Yarn and monofilament are cold-drawn or orientated to effect a very considerable increase in tensile strength. Specifications: ASTM D789.

Polydichlorostyrene. Thermoplastic prepared by polymerization of dichlorostyrene. It is similar in most respects to polystyrene except that a considerable increase in heat distortion temperature has been made with only a slight sacrifice in electrical properties. Trade name: *Styramic H.T.*

Polyethylene. Prepared by the polymerization of ethylene, polyethylene is outstanding for low electrical losses at high frequencies and has found extensive application as a dielectric in r-f coaxial cables. In thin sections it is flexible, but thick sections are rigid and can be machined. It is thermoplastic and is fabricated by injection or extrusion. It is available in the forms of molding compound, tape, tubing, monofilament, and rods or slabs. English practice includes plasticizing with polyisobutylene, which lowers the cold-brittleness point. Polyethylene is insoluble in common solvents when cold but dissolves in hot hydrocarbons. Coatings of polyethylene may be applied by flame-spraying or by the use of emulsions. Coatings have a low moisture permeability. Trade name: *Polythene.*

Polyflex.* Plax Corp. Flexible polystyrene sheet.

Polystyrene. Thermoplastic resin produced by polymerization of monomeric styrene with heat and sometimes a catalyst such as a peroxide. Polystyrene is outstanding for low dielectric loss at high frequencies. It has high dielectric strength and good arc resistance. It has zero water absorption and good mechanical strength, and it does not become more brittle at low temperatures. It has exceptional optical clarity and high refractive index. Unfortunately, it cannot be used at temperatures much above 65 deg cent without cold flow occurring, and some tendency for surface crazing exists. Crazing may be minimized by suitable heat treatment to remove surface strains. Attempts to increase the operating temperature by the use of fillers have not been too successful, for they increase the tendency to crack. The usual method of fabrication is injection molding, although many parts are machined from sheet, rod, or tube stock. Polystyrene may be drawn or oriented to form a flexible sheet (*Polyflex* *) or plasticized to make films for use in capacitors, etc. Specifications: ASTM D703. Trade names: *Loalin, Lustron, Plax, Polyflex, Styron.*

Polystyrene, Modified. Polystyrene has been combined with chlorinated diphenyl to form a noninflammable plastic with a heat distortion point somewhat higher than that of polystyrene. It is also easier to machine, but the electrical losses are slightly higher. Trade name: *Styramic.* Other modifications are possible, such as copolymerizing styrene with other materials such as butadiene. With about 25 per cent styrene **Buna** S rubber is formed, but if the styrene is considerably in excess a semi-flexible thermoplastic (*Styraloy* *) is produced. This material has good dielectric properties and is very tough. It is suitable for wire and cable insulating. *Cerex* * is another recently introduced styrene copolymer with improved heat resistance, high strength and hardness, and unusual chemical resistance, but somewhat higher losses than polystyrene.

Polytetrafluoroethylene. Manufactured by polymerizing gaseous tetrafluoroethylene; a plastic with remarkable heat, chemical, and solvent resistance. Polytetrafluoroethylene is very tough over a wide temperature range and has a loss factor less than that of polystyrene; its dielectric constant of 2.0 is the lowest of any solid insulating material. It is very expensive as of 1949, and extrusion or molding is slow. Some machining is necessary on most parts since molding is very difficult. Trade name: *Teflon.*

Polythene.* E. I. du Pont de Nemours & Co. **Polyethylene.**

Polyvinyl. See **Vinyl.**

Porcelain. A ceramic body usually composed of clay, feldspar, and flint, finely ground, mixed with water, formed to desired shape, dried, and fired at temperatures usually ranging from 1300 to 1800 deg cent. When desired, glaze is applied on all surfaces, except the base on which the part rests in the kiln during firing, by painting on a composition which fuses to a translucent or transparent glass. Mixtures of china clay or **kaolin,** which are slightly plastic, and ball clay, which is very plastic, are used to give the necessary working properties to the wet dough. Feldspar is a naturally occurring potassium aluminum silicate. Flint is added in the form of ground sand or quartz. Normal porcelains contain from 20 to 60 per cent of clay, from 15 to 50 per cent of feldspar, and from 0 to 65 per cent of flint. Magnesia sometimes is added up to 50 per cent to improve the strength at high temperatures. Special porcelains vary widely in composition. Some are made from natural aluminum silicates such as andalusite with enough clay to give a bond. Magnesium silicate ceramics made with bases of talc, **steatite,** etc., give modified porcelains with superior electrical and mechanical properties. The raw "body" or dough is formed into shape by two distinct processes. In the dry process the mass is compressed in steel dies. Parts must have "draft" and taper similar to die castings, and a tolerance of plus or minus 1/64 in. per in. is necessary to allow for shrinkage variation in firing and wear of molds by abrasion. Minimum commercial tolerance on thickness is plus or minus 0.010 in. Dry-process porcelain is used for insulation under 5000 volts only, since it is porous and weaker mechanically and electrically than wet-process porcelain. The porosity is from 3 to 5 times as high as that of wet-process parts. Wet-process porcelain is made by forming the dough to the approximate shape, drying, and machining on vertical or horizontal lathes to final shape. It also is sometimes cast in a fluid state in absorbent molds which remove enough water to enable the part to be removed after some time and dried. Cast porcelain compares favorably with formed wet-process porcelain. Wet-process porcelain has a very low porosity and high dielectric strength. A tolerance of about 1/32 in. per in., plus or minus, is necessary for commercial manufacture. All porcelain is relatively weak in tensile, flexural, and impact strength. It has poor resistance to thermal shock except in special grades. The mechanical strength depends upon the flint content, the heat resistance upon the clay, and the dielectric strength

* Trademark names.

upon the feldspar, which unfortunately tends to make the parts brittle. Porcelain is comparatively inexpensive and chemically inert. Dry-process parts are considerably cheaper than wet-process parts. Glazing improves the resistance to weathering, but it must have a coefficient of expansion similar to that of the body, for a cracked or "crazed" glaze reduces the strength. The insulation resistance of normal porcelains drops rapidly above 300 deg cent; to minimize this effect, alkali metals are reduced in amount as much as possible. For high-temperature work free quartz in the fired body is undesirable since it has irregularities in its thermal expansion curve which tend to cause cracking, and so it is eliminated as far as possible. The desirable structure is usually crystals of aluminum silicates (known as mullite and sillimanite) evenly embedded throughout a glassy matrix. Pores are, of course, highly undesirable. Fired porcelain parts can be ground to meet close tolerances, and two or more subparts can be fastened together with neat Portland cement, litharge-glycerin cement, or asphalt and resin base compounds. Low-melting metals, such as babbitt, may be cast around porcelain with some attendant danger of cracking.

Pressboard. A material similar to paper except that it is thicker, less flexible, and usually denser. Grades containing above 50 per cent cotton fiber may be formed by heat and pressure into simple shapes. The better grades are also known as fullerboard. Pressboard is much used for low-frequency coil construction with subsequent impregnation. The material, of course, is hygroscopic and must be treated with oil, wax, varnish, or other compounds to increase dielectric strength and repel moisture. The impregnated material is a cheap and satisfactory insulator where the highest dielectric properties are not required.

Prestite.* Westinghouse Electric Corp. Special dry-process **porcelain.**

Pyralin.* E. I. du Pont de Nemours & Co. **Cellulose nitrate.**

Pyrex.* Corning Glass Works. Electrical, heat, and chemical resistant **glasses.**

Pyroxylin. See **Cellulose nitrate.**

"Q" Max.* Communication Products Co. Low-loss r-f coil lacquer.

Quartz, Fused. Silicon dioxide fused at 1750 deg cent to a clear, translucent, glassy mass. Very stable, will not absorb water, and is an exceptional insulator. Strong mechanically and extremely resistant to thermal shock on account of the low coefficient of expansion. It is not attacked by solvents or solutions except by hydrofluoric acid and slowly by concentrated alkalies. For high temperatures it must be kept clean, as traces of metallic salts or oxides will flux the quartz to form a low-melting glass and cause failure of the tube or other device. It is available in rods, blocks, tubes, and extruded shapes. Special shapes can be cast in graphite molds. It can be ground and disk-sawed readily since it does not crack easily. It is very expensive.

Rayon. The three common types of rayon are acetate, viscose, and cuprammonium. Viscose rayon is the strongest and most suitable for protective braids on hook-up wire but is not so resistant to abrasion as cotton. Viscose rayon contains traces of residual sulfur which may cause corrosion if it is used for magnet wire insulation; acetate and cuprammonium rayon are suitable for such applications. Acetate rayon is especially suitable for use with very fine wires for it is non-corrosive even under conditions where electrolysis would take place.

Resimene.* Monsanto Chemical Co. **Melamine** molding compounds.

Resinox.* Monsanto Chemical Co. **Phenolic** molding compounds.

Resistoflex.* Resistoflex Corp. Polyvinyl alcohol products.

Rosin. Rosin is a natural resin obtained by steam distillation of turpentine and rosin oils from the exudations of certain varieties of pine trees. Rosin comes in letter grades. WW (water white) is the best, descending in reverse alphabetical order to A and B, which are very impure grades, containing much dirt, and almost black. It is universally graded by the color. Although the properties do not vary directly with the color, the color is important for use in varnishes. Rosin in grades from WW to H is extensively used in oil and wax impregnating compounds. It also is very cheap. Probably the most important use in the electrical industry is non-corrosive soldering flux. Solutions of rosin in No. 1 S.D. alcohol, or in alcohol and ethyl acetate, form a substantially non-corrosive soldering flux and yet the activity of rosin at the temperature of melted solder is sufficient to remove thin coats of metallic oxides and insure a good joint on tin, copper, brass, or nickel silver. It is unsatisfactory for steel. Any rosin left around the joint is non-conductive, which is a further advantage.

Rubber, Cyclicized. Thermoplastic resin derived from natural or special synthetic rubber by treatment with stannic chloride or chlorostannic acid. This resin is extremely resistant to moisture diffusion and may be added to wax mixtures to decrease cracking at low temperatures. Trade names: *Pliolite*, *Marbon B*.

Rubber, Hard. Hard rubber is usually vulcanized with 20 to 30 per cent sulfur, in the form of sheets, rods, tubes, or molded shapes. It is also called vulcanite and ebonite, and is known under various trade names. It is a hard, dense material, is easily machinable, takes a high polish, and is resistant to wear. At temperatures slightly below room temperature it becomes increasingly brittle, and at higher temperatures it softens and flows under pressure. Under heavy load, "cold flow" occurs at room temperature. In a small intermediate temperature range it is tough and almost "unbreakable." It is combustible but is not easily ignited. It has low water absorption and is immune from attack by most acid and alkali solutions and fumes. It is attacked and swelled by oils and rubber solvents. It is attacked by ozone, although less than soft rubber, but special grades are available with improved oil and ozone resistance. It is resistant to sparks but will not withstand heavy arcs. The sulfur is never fully combined, which leads to some serious difficulties. The sulfur has a tendency to appear in a surface film known as "bloom," causing discoloration. Ultraviolet light produces sulfuric acid from this layer, seriously lowering the surface resistivity and causing corrosion of nearby metals. "Blooming" can be greatly reduced by careful compounding. Metallic inserts should be protected by a coating of tin or other corrosion-resistant substance. The tendency to "cold flow" and the high co-

* Trademark names.

efficient of expansion can be reduced by incorporating suitable fillers such as talc; lower grades of hard rubber usually are filled for economic reasons, however. Rubber can be "preformed" before molding to nearly the final shape and hence may be used for fine tubes and thin-walled sections not easily obtainable with phenolic moldings, but the accuracy in molding is usually much lower, owing to shrinkage, distortion, and high coefficient of expansion. Hard rubber is easily machined by normal methods, but grinding is sometimes more economical. Special drills also give improved performance, and lubricants are valuable for drilling, tapping, and turning. Tungsten carbide and diamond tools give more nearly satisfactory production. The material may be sheared and punched if heated. Many moldings as well as machined parts require polishing with pumice on a moderately hard "buff" at low speed to avoid excessive heat. The material usually is black but can be obtained in a number of colors, mostly with high filler content. Hard rubber has very high dielectric strength and resistivity, and low dielectric constant and power factor, but all the dielectric properties are affected seriously by rising temperature. The mechanical temperature limit is about 45 deg cent for unloaded and 70 deg cent for loaded hard rubber with light pressures.

Rubber, Synthetic. The principal synthetic rubbers are **Buna S, Buna N, neoprene, butyl,** and *Thiokol,** although elastomeric vinyl compounds might also be classed as synthetic rubber. All the above are discussed elsewhere in this section.

Rubber, Vulcanized. Sulfur and rubber react at temperatures in the vicinity of 100 deg cent to form a tough, elastic, strong material. The crude rubber is washed, sheeted, and dried. The sheets are then mixed on hot rolls with sulfur, fillers, plasticizers, accelerators, and anti-oxidants as wished, and sheeted, molded, or extruded to the desired forms. The article then is vulcanized by heating to temperatures from 125 to 145 deg cent for soft rubber articles, and 160 to 170 deg cent for hard rubber parts. With plain rubber, from 2 to 10 per cent sulfur gives a soft rubber stock. Hard rubber contains from 20 to 32 per cent sulfur. With plain sulfur, vulcanization or "cure" may take 2 or 8 hours, but by means of accelerators the time may be reduced to as low as 20 minutes. Litharge, lime, and magnesia are inorganic accelerators as well as fillers. Complex organic compounds, such as tetramethylthiuram disulfide, phenylguanidines, and mercaptobenzothiazole, function to give fast cures which are not critical as to the time required to obtain maximum physical properties and are known as "flat" cures. Fillers in the form of fine powders, such as carbon black, zinc oxide, clay, and whiting, are employed in nearly all rubber goods. They cheapen the compound, of course, but they also increase the strength and toughness. Soft rubber compounds usually contain from 50 to 80 per cent of filler. Although the natural resins and proteins assist in the "milling" or breaking down of the rubber to some extent, very often improved working is obtained by adding plasticizers or softeners, including paraffin, waxes, para-cumaron resin, oils, fats, and asphaltic and bituminous materials in various percentages. Reclaimed rubber also improves the working properties. So-called mineral rubber, which is an asphaltic residue, is sometimes added up to 20 per cent and can be classed as a filler. The dielectric constant of rubber-sulfur compounds at 25 deg cent rises from 2 to 11 per cent of sulfur and falls again from 16 to 19 per cent sulfur, and then changes very slightly up to 32 per cent sulfur. The power factor goes through somewhat similar variation, starting at 8 per cent sulfur. The maxima of these curves are displaced to higher sulfur content by increasing temperature. The normal soft and hard rubber compositions thus have low dielectric constant and power factor, and the intermediate region, which is seldom used, has poor electrical properties. Resistivity rises in a fairly regular curve from 2 to 28 per cent sulfur. (See *Bureau of Standards Scientific Paper* 560, part II, by H. L. Curtis, A. T. McPherson, and A. H. Scott.) Softeners change the dielectric constant slightly but may seriously increase the power factor in quantities of only 10 per cent. The dielectric constant is increased nearly proportionally to the filler content. Carbon black causes a sharp increase of dielectric constant from 2.7 to 6.0. Zinc oxide, lead oxide, and selenium show much slower rates of increase. Powdered quartz gives only a slight increase. The effect on the power factor is much the same: 20 per cent carbon black elevates the power factor from 0.0025 to nearly 0.05. Increasing quartz content slightly improves the power factor. The introduction of carbon black greatly reduces the resistivity of rubber. Carbon black is, however, the best filler from a mechanical standpoint: a several-fold increase in the tensile strength is produced. Rubber compounds absorb water, which causes an increase of dielectric constant and power factor, and decrease of resistivity, but on long immersion the power factor may decrease again. Water absorption can be considerably lowered by extended washing of the crude rubber to remove water-soluble matter and by the use of water-insoluble fillers such as silica, zinc oxide, or hard rubber dust. Absorption of water is less in sea water than in distilled water. The usual grade of wire insulation contains 20 per cent minimum of rubber; better grades have 30 per cent, and high grades 40 per cent. Covering of portable cords, etc., subject to mechanical wear may contain up to 60 per cent. The mechanical and dielectric strength of rubber is lowered by the action of oxygen and more rapidly by ozone which is present in corona discharge: the rubber cracks when normally vulcanized and may melt if the temperature is high and the compound is undervulcanized. In order to improve the resistance to oxygen, organic compounds, such as diphenylamines or hydroquinone, known as antioxidants, are added in small percentage. Rubber stocks are tested for this defect under 300 lb per sq in. in oxygen gas at 70 deg cent and by contact with ozone at atmospheric pressure. Stretching the rubber under test seriously increases the rate of ozone attack. Ultraviolet light also sharply accelerates the combination with oxygen. Heat deteriorates rubber rapidly; 49 deg cent is the maximum operating temperature for Code rubber insulation. Performance grades are satisfactory at 60 deg cent, and heat-resistant grades at 75 deg cent. Rubber is swelled quickly and eventually dissolved by hydrocarbon solvents and oils. Special grades are available which minimize this defect. At low temperatures a low-sulfur rubber compound is no longer elastic, and a piece stretched and cooled to −20 deg cent or lower will not return to its original length until the temperature rises. The power factor also increases sharply to a maximum in this region, with a value over 10 times the value at 20 deg cent,

* Trademark names.

indicating a change of state in the compound. The power factor also rises with increasing temperature, but less rapidly.

Rutile (TiO_2). A particular crystalline form of titanium dioxide, with a dielectric constant of 80 to 110, which is used in the manufacture of high-dielectric-constant ceramics.

Saflex.* Monsanto Chemical Co. Polyvinyl butyral.

Saran.* Dow Chemical Co. Vinylidene chloride.

Saturated Sleeving. Cotton sleeving impregnated with thin varnish or compound which does not fill completely the interstices of the fabric. Dielectric strength is low, and resistance to humidity is poor. It is valuable mainly to space conductors apart.

Scotch Tape.* Minnesota Mining and Mfg. Co. Pressure-sensitive, non-corrosive electrical tape with various backing materials.

Shellac. Produced by the insect *Tachardia lacca*, which attaches itself to numerous species of trees native to India, Indo-China, and Siam, and excretes the resin at several parts of the 6-month life cycle. The kusmi tree is the most valuable host to the parasite lac insect since the lac therefrom is of higher quality. Crude lac consists of about 75 to 85 per cent resin, 2 to 4 per cent dye, 1 to 2 per cent ash, 2 to 3 per cent water, and 9 to 15 per cent residue and dirt. The natives wash the crude lac, dry it, and press it through a bag with the aid of heat. The mass is plastered into a sheet on a heated object and stretched while still hot. After it cools it is broken into flakes and shipped in bags. Native shellac is sometimes adulterated with rosin. Machine-made shellac is extracted by various processes employing solvents or heat. Refined shellac contains over 90 per cent resin, from 3 to 5 per cent of wax, 1 to 2 per cent moisture, and 1 to 5 per cent of matter insoluble in alcohol. Shellac is sold in many grades which are principally determined by the color. D.C. is a very high grade, free from dirt. Superfine is made from best kusmi lac. Fine and Standard No. 1, T.N., and garnet lac follow in descending order of quality. Rosin is limited by trade standards to 3 per cent maximum except in garnet shellacs which may sometimes contain 20 per cent. Machine-made shellac is graded somewhat differently. One manufacturer's grades are Fine, Superfine, and ABTN. All are hard, pure shellacs. BB is used for blending, and amber where a light color is required. Completing the list are: T.N.S. pure orange shellac, various grades of T.N. shellac, and garnet. Machine-made garnet is low in wax and rosin, and is valuable for insulating use. Shellac is naturally variable in quality, depending on the source and methods of collection. A small amount of orpiment (arsenic sulfide) is added to some orange shellac to lighten the color. Orpiment is insoluble in solvents, but otherwise usually has no beneficial or harmful effect. Shellac has been used for some time in hot molding compounds, in insulating varnishes, and as a binder for composite insulations of paper, mica, etc. Various mineral fillers, such as asbestos, powdered mica, clays, marble, and wood flours, are blended with shellac on rolls in a manner similar to rubber compounding. The material is sheeted off the rolls, and blanks are cut to size. The preheated blanks are usually molded in steam-heated molds at about 160 deg cent under hydraulic pressure for about 1 minute, the dies are then chilled with water, and the piece is removed. Little or no chemical action takes place. The finish is excellent if the mold surface is polished, but parts can be polished subsequently. Accuracy of molding is about the same as that of phenolic moldings, and the dies are much the same except for the cooling feature. The moldings have good weather resistance and are fairly impervious to moisture, but they are somewhat brittle at low temperatures and soften at 75 deg cent. It is not well known that shellac is somewhat thermosetting under certain circumstances. Long-continued heating above 100 deg cent will solidify the melted shellac to a tough, horny mass. This action is accelerated by increased heat and by hexamethylenetetramine, aluminum chloride, urea, and other agents, and is retarded by alkalies, alkaline salts, aniline, and other substances. (See *Bulletin* 14, Indian Lac Research Institute.) Shellac is soluble in alcohols and ketones, and the solutions are used for varnishes and cements. Shellac loses solubility on standing for long periods of time, and the plasticity is also decreased somewhat. The flexibility of shellac films may be increased by plasticizing with castor oil or tricresyl phosphate.

Silaneal.* Dow Corning Corp. Silicone treating fluid for ceramics.

Silastic.* Dow Corning Corp. Silicone rubbers.

Silica, Fused. Fused silica is similar to translucent fused quartz in its properties, but it is not made from as pure a sand and usually contains some iron. It is an excellent insulator.

Silicones. Silicones are a class of organo-silicon compounds with a chemical structure analogous to that of hydrocarbons, but with the carbon atoms of the chain replaced by silicon atoms with an oxygen atom inserted in each bond between the silicon atoms. By attaching various hydrocarbon chains to each silicon atom, and varying the chain length by different polymerization procedures, fluids, greases, plastics, and resins are produced. The general properties of silicones compared to those of hydrocarbons are: (1) improved heat resistance; (2) smaller change in viscosity with temperature; (3) resistance to oxidation; (4) resistance to arcing; (5) high flash and fire points. In common with some hydrocarbons, they have low power factors and are water-repellent to a high degree. Silicone fluids are available in viscosities from 0.65 to 1000 centistokes at 25 deg cent, and in volatilities from nearly zero to approximately that of water. They have dielectric constants of 2.4 to 2.75, and power factors of 0.0002 from 100 cycles to 10 Mc, rising to 0.0006 at 100 Mc. Power factor increases with temperature but is always less than that of a good grade of mineral oil. Silicone fluids may be used to treat glass or ceramics (*Silaneal;* Dri-film* *) to produce a water-repellent surface, giving a greatly increased surface resistivity under condensation conditions. Silicone greases (DC No. 4 Ignition Sealing Compound) may be used to fill connectors to prevent arc-over, corona, or leakage, or to render surfaces water-repellent. Silicone resins are used in conjunction with glass textiles to form heat-resistant boards, cloths, and wire insulation, and in the form of varnishes for coil impregnation. Silicone rubber (*Silastic* *) is heat-resistant to 250 deg cent, remains flexible down to -55 deg cent, and has good dielectric properties.

---

* Trademark names.

Bass, S. L., and T. A. Kaupp, *Proc. I.R.E.*, Vol. 33, 441 (July 1945).
Johannson, O. K., and J. J. Torok, *Proc. I.R.E.*, Waves and Electrons section, Vol. 34, 296 (May 1946).
Norton, F. J., *Gen. Elec. Rev.*, August 1944.

Silk. Silk is obtained from cocoons spun from double continuous filaments secreted by the "silkworm," which is the larva of the *Bombyx mori* and other moths. The fiber is unwound from the cocoon by hand, usually scoured (degummed) to remove the natural sticky gum or wax, called sericin, which cements the duplex filaments, and twisted into thread. Cultivated silkworms are fed on mulberry leaves; wild silkworms give a coarser quality known as tussah silk. Orgazine silk is from the best selected cocoons, and tram silk is from the poorer cocoons. Floss silk is spun from broken lengths of filaments. Silk for insulation should be free of loading materials and as well washed as possible, since thorough washing greatly improves the insulating qualities. Silk is used in the form of cloth and tapes, plain or varnished, and as wrapped or braided insulation on wires. Silk insulation for a given dielectric strength is thinner than cotton and has somewhat higher insulation resistance and lower dielectric constant but is not as resistant to heat. Silk flock is sometimes used to add strength to molding compounds.

Sisal Hemp. A bast cordage fiber obtained from the leaves of the century plant or agave. In strength and length of fiber it is inferior to manila hemp. It is used to some extent in making paper and pressboard, and for reinforcing large molded laminated parts.

Spauldite.* Spaulding Fibre Co. Phenolic laminates.

Steatite. Principal ceramic used for radio apparatus. Made chiefly from magnesium silicate which, after firing, forms clinoenstatite crystals. Low water absorption and excellent electrical properties are characteristic. Special grades (L-5) are available with even lower losses than standard or regular grades (L-4 or L-3). Specifications: Joint Army-Navy Spec. JAN-I-10, Grades L-3, L-4, or L-5.

Styraloy.* Dow Chemical Co. Elastomeric polystyrene copolymer.

Styramic.* Monsanto Chemical Co. Polystyrene and chlorinated diphenyl molding compound.

Styramic H. T.* Monsanto Chemical Co. Polydichlorostyrene molding compound.

Styrene. Volatile liquid monomer, also known as vinyl benzene, used for manufacture of polystyrene, Buna S rubber, and other plastics. Polymerizes spontaneously in time to polystyrene or more rapidly with the aid of heat or a catalyst. It also is used as a fully reactive constituent in polyester laminating liquids so that no solvent evaporation is necessary.

Styrofoam.* Dow Chemical Co. Expanded polystyrene.

Styron.* Dow Chemical Co. Polystyrene of various types.

Synthane.* Synthane Corp. Phenolic laminates.

Teflon.* E. I. du Pont de Nemours & Co. Polytetrafluoroethylene products.

Tegit.* Garfield Mfg. Co. Asbestos coal-tar moldings.

Tenite I.* Tennessee Eastman Corp. Cellulose acetate.

Tenite II.* Tennessee Eastman Corp. Cellulose acetate-butyrate.

Textolite.* General Electric Co. Phenolic and other molded or laminated products.

Thalid.* Monsanto Chemical Co. Low-pressure or contact laminating resins.

Thiokol.* Thiokol Corp. Polysulfide rubbers.

Transite.* Johns-Manville Corp. Portland cement and asbestos molded products.

Tuf-flex.* Libby-Owens-Ford Glass Co. Tempered glass.

Turx.* International Products Corp. Glass-bonded mica.

Urea Resins. Reaction of urea, $CO(NH_2)_2$, and formaldehyde produces methylol ureas which are water-soluble. Paper, alphacellulose, cloth, or wood is impregnated with solutions and cured with heat and catalysts to a thermosetting, water-insoluble plastic. Urea moldings are light weight, rigid, and hard. They have high dielectric strength, good arc resistance, and moderate electrical losses. Impact strength is lower than that of phenolic materials. Translucent moldings in any color may be obtained. Specifications: ASTM D705. Trade names: *Beetle, Plaskon.*

Varnishes, Insulating. Varnishes are generally classified according to composition, as oleoresinous or "oil varnishes" and "spirit varnishes," but some commercial types do not fall strictly into either class. Oil varnishes are made by combining a resin, commonly a copal, with a drying oil. It is usually necessary to melt or "run" the resin before adding the oil to get a clear solution. Varnishes with a high oil content are known as "long-oil" varnishes; they are slow drying but deposit very flexible films. "Short-oil" varnishes have high resin content and deposit a hard film. Oil varnishes harden as a result of oxidation and polymerization of the drying oils, such as linseed, tung (China wood), or soya bean, and often of the resin or asphalt as well. Drying is accelerated by adding small percentages of catalytic agents, called "driers," usually in the form of resinates or linoleates of cobalt, lead, or manganese, which increase the rate of oxidation. Short-oil varnishes will air-dry in 4 to 18 hr, depending on the type, to a reasonable degree of hardness. Long-oil varnishes will not air-dry in a reasonable time. By baking at 100 to 110 deg cent the drying time can be shortened to 2 to 8 hr because of the faster oxidation. Baking produces a harder film and gives better adhesion to the object. It also serves to drive out moisture from fibrous materials which are being treated, and to improve the dielectric properties. Many modern insulating varnishes contain synthetic resins of the phenol-aldehyde or alkyd type which give hard durable films as the result of the thermosetting of the resins during baking. Thermosetting resins are also used dissolved in "spirit"-type solvents instead of oils, yielding a moderately hard film, on air-drying, which is increased in hardness and durability by baking. The dielectric properties of the synthetic resin varnishes are usually excellent. Asphalts are used with resins in some black oil varnishes and without oil or resin in the so-called asphaltum varnishes which are merely solutions of asphalts in benzine or other hydrocarbons. These varnishes dry principally by evaporation, but the last traces of solvent leave the asphalt very slowly, and if the object is heated to drive off the solvent many of the asphalts will oxidize and polymerize to a certain extent, producing a fairly hard film.

* Trademark names.

Spirit varnishes dry by evaporation of the solvent, leaving a film of the dissolved resin, asphalt, or gum. Cellulose nitrate and acetate varnishes are in this class but are usually considered separately as "lacquers." Many of the resins leave a brittle film when used alone so that a soft gum or plasticizing agent like castor oil is usually added to give the necessary flexibility, except in shellac which ordinarily does not need plasticizing. The principal spirit varnish resins are shellac, manila copal, dammar, mastic, kauri, and sandarac. The solvents and thinners used are alcohols, esters, hydrocarbons, and turpentine. Spirit varnishes are commonly air-drying, although the evaporation of the solvent is often hastened by moderate heating. Spirit varnishes are generally not used for impregnation but are common for external coats and for sticking, and for bonding of mica and other materials. For external coating work, varnishes are applied by spraying, dipping, roller coating, or brushing, depending on the nature of the work. For impregnation, dried articles are dipped while still hot into a varnish of low viscosity and low surface tension to insure thorough penetration. Dipping time must be determined for each article by experiment. A much better impregnation is secured by drying coils in a vacuum and admitting the varnish to the work container, then "breaking" the vacuum and applying pressure. Impregnated coils should be drained and baked at 100 to 110 deg cent for a period sufficient to harden the varnish film. Higher temperature tends to disintegrate fibrous materials in prolonged baking, and lower temperatures do not remove moisture. For short drying schedules, temperatures up to 150 deg cent are sometimes employed satisfactorily. Objects should be exposed to fresh currents of air during drying in order to remove solvent vapors which retard hardening. For methods of testing dielectric strength, heat endurance, and oil proofness, see ASTM D115.

Varnished Cloth. A suitable fabric coated with yellow or black insulating varnish so that the fibers are thoroughly impregnated. Varnished silk usually is from 0.003 to 0.005 in. and cotton from 0.005 to 0.040 in. thick. Tensile strength of cotton-base cloth per inch width of warp runs from 45 to 100 lb, and Elmendorf tearing strength across the warp varies from 100 to 300 grams. Dielectric strength usually runs from 800 to 1500 volts per mil.

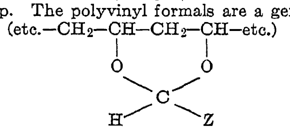
Varnished Tubing. Commonly called "spaghetti," magneto tubing, etc., according to manufacturer and grade; made by coating, or impregnating and coating, cotton sleeving with varnishes. ASTM Specification D372 distinguishes three grades (see specification for details). Grade A is generally known as flexible varnished tubing, or motor and transformer tubing, or impregnated magneto tubing; it has a minimum dielectric strength of 7000 volts average. Grade B, generally known as "radio spaghetti," has a minimum average dielectric strength of 4000 volts. Grade C is similar to saturated sleeving, and its dielectric strength is lower.

Vibrin.* Naugatuck Chemical Co., Div. U. S. Rubber Co. Liquid polyesters and cross-linking monomers.

Vinyl Plastics. An extremely important class of thermoplastic resins, composed of linear chains formed by the polymerization of monomers of the general type $\left(CH_2{:}C\begin{smallmatrix}X\\Y\end{smallmatrix}\right)$. The result is a continuous chain $\left(\text{etc.—}CH_2\text{—}\overset{X}{\underset{Y}{C}}\text{—}CH_2\text{—}\overset{X}{\underset{Y}{C}}\text{—etc.}\right)$.

If group X is	*and group Y is*	*the product is*
hydrogen	hydrogen	polyethylene (*Polythene* *)
hydrogen	chlorine	polyvinyl chloride (*Geon* *)
chlorine	chlorine	polyvinylidene chloride (*Saran* *)
hydrogen	hydroxy (—OH)	polyvinyl alcohol
hydrogen	phenyl (C_6H_5—)	polystyrene
hydrogen	carbazyl ($C_{12}H_8N$—)	polyvinyl carbazole (*Polectron* *)
hydrogen	acetoxy ($CH_3\cdot CO\cdot O$—)	polyvinyl acetate
methyl (CH_3)	methylcarboxy ($CH_3O\cdot OC$—)	methyl methacrylate (*Plexiglas*,* *Lucite* *)

If mixtures of two monomers are copolymerized, a copolymer such as polyvinyl chloride-acetate is produced in which all the Y's are hydrogen, most of the X's are chlorine, and the remainder of the X's are the acetoxy group. The polyvinyl formals are a general group of polymers with modified chains constructed like this:



If group Z is	*the product is*
hydrogen	polyvinyl formal (*Formex*,* *Formvar* *)
methyl (CH_3—)	polyvinyl acetal (*Alvar* *)
butyl (C_3H_7—)	polyvinyl butyral (*Butacite*,* *Butvar*,* *Saflex*,* *Vinylite** X)

Of course, different conditions of temperature, pressure, catalyst, and carrier solvent or emulsion are required for each case, and the length of the chain of molecules is varied to suit requirements by altering these conditions. Only those plastics with vinyl in the product name are usually classed as vinyl plastics, but by this illustration the relationship of many of the thermoplastics is immediately apparent.

Vinyl Chloride. Perhaps the most important vinyl plastic from a tonnage viewpoint, vinyl chloride is very extensively used for hook-up wire insulation, cable jackets, insulating tubing, and tape. It is

* Trademark names.

also employed for non-rigid or elastomeric molded parts since it has rubberlike characteristics when plasticized properly. The amount and kind of plasticizer may be varied over wide limits to form compounds ranging from hard, stiff, non-extensible plastics, to limp, flexible, and high-elongation products resembling rubber. The electrical properties and inflammability likewise vary with the kind and amount of plasticizer, but, in general, these resins have a rather high dielectric loss and a lower insulation resistance than those of rubber compounds, although they have a high dielectric strength. Wire insulation for operating temperatures of 80 deg cent is now available, although 60 deg cent has been the limiting temperature for most older compositions. These compounds are nearly non-inflammable and do not oxidize like rubber but are subject to some stiffening due to loss of plasticizer with time. Except for flexing at low temperature the life should be very long. Specifications for resin: ASTM D728. Trade names: *Flamenol*, *Geon* 100 series, *Vinylite* Q series, *Koroseal.*

Vinyl Chloride-acetate. This resin is similar to straight polyvinyl chloride resin except that about 5 per cent of the resin may be vinyl acetate, which improves the processing and extrusion characteristics and reduces the amount of plasticizer required. Uses and characteristics are similar to those of polyvinyl chloride, but in addition these resins are used for rigid, transparent, or colored sheets of high dimensional stability and good electrical properties. These resins also have proved to be very successful for high-quality phonograph records and for manufacture of textiles (*Vinyon*). Specifications for rigid sheet: ASTM D708; for non-rigid resin: ASTM D742. Trade names· *Vinylite* V series.

Vinylidene Chloride. Similar to vinyl chloride resins but contains more chlorine, is harder, and is capable of orientation by drawing or other cold work to increase the tensile strength. For molding purposes, it is commonly copolymerized with 10 per cent of vinyl chloride to serve as an internal plasticizer. Though the electrical properties are not outstanding, the strength, toughness, and non-inflammability of this plastic have made it valuable for fungus-resistant cable braids and ropes, as tubing for water cooling, etc. Specifications for molding compounds: ASTM D729. Trade name: *Saran.*

Vinylite.* Carbide and Carbon Chemical Co. Vinyl resins, of which various grades are distinguished by suffix letters.

Vinyon.* Carbide and Carbon Chemical Co. Polyvinyl chloride-acetate textiles.

Vistanex.* Standard Oil Co. of N. J. Polyisobutylene.

Vitreosil.* Thermal Syndicate, Inc. Fused silica.

Voltron.* Industrial Synthetics Corp. Vinyl tubing and tape.

Vulcabeston.* Johns-Manville Corp. Asbestos with rubber or gum binder.

Vulcoid.* Continental-Diamond Fibre Co. Resin-impregnated vulcanized fiber.

Vycor.* Corning Glass Works. High-silica glass.

Waxes. Waxes are of three origins: animal, vegetable, or mineral. The principal animal waxes are: beeswax, m.p. 63 deg cent; wool wax, m.p. 35 deg cent; spermaceti, m.p. 49 deg cent; insect or Chinese wax, m.p. 81 deg cent. The principal vegetable waxes are: carnauba, m.p. 85 deg cent, and candelilla, m.p. 68 deg cent. The principal mineral waxes are: montan, m.p. 72 deg cent; ozokerite refined to form ceresin, m.p. 65–73 deg cent; paraffin waxes obtained from petroleum. Waxes in general are not "wetted" by water and are very resistant to penetration by moisture. Ceresin and non-crystallizable high-melting-point paraffin waxes such as *Superla,** *Syncera,** and *Cerese wax* * have higher insulation and moisture resistance than the low-melting paraffins and the animal and vegetable waxes. Waxes which crystallize, such as montan, have a tendency to crack and admit moisture by capillary action; mixture with a soft wax or resin usually minimizes this tendency. Wax compounds are hardened and the flow point raised by incorporating rosin or other resins and higher-melting waxes such as ceresin, montan, or carnauba. Most naturally occurring waxes have low dielectric constants and low power factors as well as high resistivity. Synthetic waxes made by chlorinating naphthalene or paraffin have somewhat higher dielectric constants and losses. Long heating of liquid waxes at high temperatures in contact with air tends to cause decomposition and development of acidity, particularly in the presence of copper, which acts as a catalytic agent to increase oxidation markedly.

Zircon Porcelain. This material consists chiefly of zirconium silicate with addition of small amounts of clay and metallic oxides to aid in manufacture. It has higher mechanical strength than other electrical ceramics, and a lower coefficient of expansion than all but cordierite. It is extremely resistant to thermal shock. The electrical characteristics at radio frequencies are excellent, except that the dielectric constant is 50 per cent higher than that of steatite and changes more rapidly with temperature. The resistivity at high temperatures is about the same as that of steatite and is superior to that of cordierite and high-voltage porcelain.

6. LIQUID DIELECTRICS

Dielectric Constant. The dielectric constant of liquids ranges from about 1.5 to almost 100. Non-polar liquids at room temperature have constants of 1.84 to 2.3, which are nearly equal to the square of the index of refraction. The dielectric constants of non-polar liquids are independent of frequency and vary only slightly with temperature, owing to thermal expansion. On the other hand, polar liquids have higher dielectric constants which vary to a marked degree with temperature and frequency in certain regions. When a polar liquid freezes, part of the influence of the dipoles is lost and the dielectric constant drops very sharply. High rates of change in dielectric constant with frequency or temperature are normally associated with high power factors at the same frequency and temperature. The dielectric constants and dipole moments of some liquids are shown in Table 2.

* Trademark names.

Table 2. Dielectric Properties of Liquids

Substance	Temperature, deg cent	Wave-length, cm	Dielectric Constant (K)	Temperature Coefficient of K, $10^{-4}/°C$ (negative)	Polar Moment, 10^{-18} esu	Conductivity, mhos/cm
Acetone	−80	∞	33.8	—	—	—
"	0	∞	26.6	—	—	—
"	+15	1200	21.85	—	—	—
"	17	73	20.7	—	—	—
"	25	∞	21.3	—	2.75	6×10^{-8}
Air (liquid)	−191	∞	1.43	—	0	—
Amyl acetate	19	∞	4.81	24	1.91	—
Amyl alcohol	20	∞	16.0	—	1.7	—
" "	18	200	10.8	—	—	—
" "	18	73	4.7	—	—	—
Aniline	18	∞	7.32	35	1.56	2.4×10^{-8}
Benzene	20	∞	2.282	8.6	0	7.6×10^{-18}
n-Butyl alcohol	20	∞	17.4	76	1.65	—
Butyl stearate	25	∞	3.3	—	—	—
Carbon dioxide (liquid)	−5	∞	1.60	—	0	—
Carbon tetrachloride	20	∞	2.3	12.5	0	4×10^{-18}
Castor oil	11	∞	4.67	107	+	1.7×10^{-11}
Chlorinated diphenyl						
mobile	25	∞	5.8	163	—	—
viscous	25	∞	5.05	133	—	—
Chlorobenzene	20	∞	5.72	31.0	1.56	—
Chloroform	20	∞	4.84	39.0	1.1	$<2 \times 10^{-8}$
Cumene	22	∞	2.2	—	0.4	—
Cyclohexane	20	∞	2.41	8.6	0	—
Cymene	22	∞	2.5	—	?	$<2 \times 10^{-8}$
Decahydronaphthalene	20	∞	2.26	6.6	0	—
Decane	20	∞	1.991	6.7	0	—
Decylene	17	∞	2.211	—	—	—
o-Dichlorobenzene	20	∞	10.2	45	2.25	—
m-Dichlorobenzene	20	∞	5.11	28.0	1.48	—
Dodecane	20	∞	2.017	7.4	0	—
Ethyl abietate	20	∞	3.95	ca. 20	+	—
Ethyl acetate	20	∞	6.15	—	1.85	$<1 \times 10^{-9}$
Ethyl alcohol	Frozen	∞	2.7	—	—	—
" "	−120	∞	54.6	—	—	—
" "	−80	∞	44.3	—	—	—
" "	−40	∞	35.3	—	—	—
" "	0	∞	28.4	—	—	—
" "	+20	∞	25.8	63.0	1.68	1.3×10^{-9}
" "	17	200	24.4	—	—	—
" "	17	75	23.0	—	—	—
" "	17	53	20.6	—	—	—
" "	17	4	8.8	—	—	—
" "	17	0.4	5.0	—	—	—
Ethyl benzene	22	∞	2.2	—	0.5	—
Ethyl ether	20	∞	4.4	46	1.15	$<4 \times 10^{-13}$
Ethylene glycol	20	∞	38.8	—	2.28	3×10^{-7}
Glycerin	25	∞	43.0	ca. 52	+	6.4×10^{-8}
"	15	1200	56.2	—	+	—
"	15	200	39.1	—	—	—
"	15	75	25.4	—	—	—
"	—	8.5	4.4	—	—	—
"	—	0.4	2.6	—	—	—
Heptane	20	∞	1.926	8.6	0	4×10^{-13}
Hexane	20	∞	1.890	9.0	0	4×10^{-18}
Hydrogen (liquid)	−258.4	∞	1.241	—	0	—
Kerosene	25	∞	ca. 2.1	—	?	$<1.7 \times 10^{-8}$
Mesitylene	22	∞	2.2	—	0	—
Methyl alcohol	Frozen	∞	3.07	—	—	—
" "	−100	∞	58.0	—	—	—
" "	−50	∞	45.3	—	—	—
" "	0	∞	35.0	—	—	—
" "	+20	∞	31.2	57	1.68	4.4×10^{-7}
Mineral oil	20	∞	2.191	4.7	0	1×10^{-16}

Table 2. Dielectric Properties of Liquids—*Continued*

Substance	Temperature, deg cent	Wave-length, cm	Dielectric Constant (K)	Temperature Coefficient of K, $10^{-4}/°C$ (negative)	Polar Moment, 10^{-18} esu	Conductivity, mhos/cm
Nitrobenzene	−30	∞	3.1		—	—
"	−13	∞	3.2		—	—
"	−5	∞	3.4		—	—
"	−4	∞	3.8		—	—
"	+15	∞	37.8		—	—
"	18	∞	36.45		3.9	2×10^{-8}
"	30	∞	35.1		—	—
Nitrogen (liquid)	−208	∞	1.44	—	0	—
Octane	20	∞	1.949	7.6	0	—
Olive oil	20	∞	3.11	36	—	2×10^{-13}
Peanut oil	11.4	∞	3.03	—	—	—
Pentane	20	∞	1.845	9.6	0	$<2 \times 10^{-10}$
Petroleum	—	2000	2.13	—	—	3×10^{-13}
Petroleum ether	20	∞	1.92	—	0	—
Phenol	45	∞	10.3	—	1.73	$<1.7 \times 10^{-8}$
n-Propyl alcohol	20	∞	22.2	53	1.66	5×10^{-8}
Pyridine	22	∞	13.9	—	2.1	5.3×10^{-8}
Quinoline	22	∞	9.0	—	2.25	2.2×10^{-8}
Rosin oil	20	∞	2.55–2.8	ca. 22	—	—
Silicone fluids						
DC 200, 200 centistokes	20	∞	2.76	34	0	10^{-14}
DC 500, 20 centistokes	20	∞	2.71	31	0	10^{-14}
Toluene	−83	∞	2.51	—	—	—
"	+16	∞	2.33	—	—	—
"	19	73	2.31	9.8	0.52	$<1 \times 10^{-14}$
Turpentine	20	∞	2.23	—	—	2×10^{-13}
m-Xylene	18	∞	2.376	8.2	0.4	$<1 \times 10^{-15}$
" "	17	73	2.37	—	—	—
p-Xylene	20	∞	2.25	—	0	—
Water (pure), frozen	−18	5000	3.16	—	—	—
" " "	−5	1200	2.85	—	—	1.6×10^{-9}
" " liquid	+17	200	80.6	—	—	—
" " "	17	74	81.7	—	—	—
" " "	17	38	83.6	—	—	—
" " "	18	∞	81.07	—	—	4×10^{-8}
" " "	50	—	—	—	—	1.7×10^{-7}

Notes. A wavelength greater than 10,000 cm is denoted by ∞.
Zero indicates that substance has no polar moment.
Plus sign indicates that substance is polar but value for polar moment was not available.

D-c Conductivity and Resistivity. The d-c conductivity of liquids is ionic in nature and has a high positive temperature coefficient. Change in conductivity with temperature is expressed by

$$G = G_0 e^{a/T} \quad \text{or} \quad G = G_0 e^{bt}$$

where G_0, a, and b are constants, T is the absolute temperature, and t is the temperature in degrees centigrade for small temperature differences. Thus a plot of the logarithm of either conductance or resistivity against $1/T$ is a straight line, and against t is approximately straight for small temperature intervals. The increase in conductivity with temperature is the result of an increase in ionic mobility arising from the reduction in viscosity. Log resistance-temperature curves therefore change slope at regions where the viscosity varies sharply, as at freezing or transition points.

The conductivity of pure liquids may be increased enormously by small amounts of impurities or moisture which readily ionize in the particular liquid. Fortunately the degree of ionization is a function of the dielectric constant, so that the non-polar liquids having a low dielectric constant are less sensitive to impurities, especially in low concentration. The difficulty of preventing contamination of liquids of higher dielectric constant has effectively prevented their use for capacitors. The resistivities characteristic of commercially pure liquids are approximately in inverse relationship to the dielectric constants, as is shown by the following table taken from The Properties of Dielectrics, by F. M. Clark, in the *J. Franklin Inst.*, Vol. 208, 17 (July 1929).

Table 3. Relation between Dielectric Constant and Resistivity

Material	Dielectric Constant	Characteristic Resistivity, ohm-cm	Material	Dielectric Constant	Characteristic Resistivity, ohm-cm
Benzene.......	2.15	4.7×10^{12}	China wood oil...	3.5	0.08×10^{12} (100° C)
Petroleum oils..	2.2	10.0×10^{12} (100° C)	Castor oil.......	4.3	0.06×10^{12} "
Paraffin wax...	2.25	5.0×10^{12} "	Ethyl alcohol....	25.0	0.3×10^{6} (18° C)
Cottonseed oil..	2.9	0.2×10^{12} "	Methyl alcohol...	31.0	0.14×10^{6} "
Asphalt........	3.1	1.0×10^{12} "	Water..........	81.07	0.5×10^{6} "
Linseed oil.....	3.3	0.61×10^{12} "			

At very high voltage gradients, some evidence of a saturation range similar to that of gases has been obtained. With long applications of voltage, impure liquids undergo a so-called electric cleaning. This is due to a very low rate of ion production so that all ions are swept to the electrodes and the conductivity drops nearly to that of a pure liquid. Solid phases also are removed in some cases by cataphoresis, but the breakdown strength is usually affected to a greater extent than the conductivity.

Dielectric Absorption and Losses. Liquids exhibit some of the phenomena of dielectric absorption shown by solids, but the rate of decrease of the initial current with time is much faster and normally is detectable only with an oscillograph. The characteristic "bound charge" of solids also is absent: the discharge current shows practically no evidence of absorption, except in highly viscous liquids of a mixed nature.

The initial high current is reduced, owing to the accumulation of space charges in front of the electrodes. The resultant non-linear potential distribution can be measured with probes; or porous cells can be used to remove the space charges from the liquid for measurement. Although the absorption results in a higher a-c conductivity, the effect is of much smaller magnitude than that in solids, and the a-c losses are much lower. The non-existence of any discharge phenomena has been attributed to the absorption of space charges by part of the electrode charges.

The power factor of most commercial non-polar insulating liquids is low, ranging from 0.0001 to 0.01. The power factor at 60 cycles is influenced by the d-c conductivity and may be expected to double for each 10 to 20 deg cent rise in temperature. At high frequencies, little change in power factor with temperature is to be expected with true non-polar dielectrics. Many mineral oils have some polar impurities which produce characteristic peaks in the power-factor curves at frequencies or temperatures where rapid changes in dielectric constant are occurring.

For an excellent summary of the effects to be expected in polar substances, consult Dielectric Properties of Organic Compounds, by S. O. Morgan and W. A. Yager, *Industrial and Engineering Chemistry*, Vol. 32, 1519 (November 1940).

Dielectric Strength. Breakdown in liquids, like that in gases, is not permanent, nor is the subsequent breakdown voltage necessarily reduced. Discussion of breakdown must be considered for two cases: pure liquids containing no dissolved or suspended gas, solid, or foreign liquid; and impure liquids.

Breakdown in pure liquids probably occurs by an ionization process similar to that in gases and undoubtedly is aided by intense voltage gradients built up by space charges near the electrodes. Change in pressure has practically no effect, but increase in temperature decreases the breakdown strength, particularly when the boiling point is approached. The time involved in the breakdown process may be as short as 10^{-7} sec with sufficient overvoltage

Impure liquids usually break down at much lower voltages for a variety of reasons, the most important of which is the presence of moisture or gases. The electric field tends to liberate dissolved gases, and, since the gas dielectric strength is only about one-tenth that of the liquid, the gas ionizes and starts the discharge. The harmful effects of moisture and gases are greatly increased by fibers or other suspended solid particles which absorb the impurity. Fibers, particles, or foreign liquids may form "bridges" or "chains" if the dielectric constant is higher than that of the liquid. Sometimes these bridges lead only to preliminary or "pilot" sparks which exert no effect on the breakdown. A very small percentage of impurity usually produces a marked lowering of breakdown strength, but larger percentages have only a slightly greater effect. Particles of carbon formed by arcs or sparks give but a small decrease, which is proportional to the concentration.

Although the breakdown voltage of pure liquids is linear with distance in uniform fields, that of impure liquids is considerably influenced by gap geometry. A horizontal gap gives lower breakdowns with impure oils than a vertical gap because of the difference in the ease of gas elimination. Small gaps are quite liable to breakdown by fiber bridges; long needle gaps are scarcely affected by most impurities. Increased gap area obviously results in lower average breakdown with impure liquids.

The effect of temperature on breakdown of impure liquids depends on the kind of impurity. Materials of low dielectric constant are more readily expelled from the field as the viscosity is reduced by raising the temperature, but the conductivity of the liquid is increased. Moisture may be expelled, raising the breakdown. Pressure increases the dielectric strength by preventing gas elimination or vaporization of the liquid.

Since formation of fiber bridges and gas elimination take an appreciable time, the strength of liquids to transient voltages is little influenced by these impurities. With steady-state currents, the strength increases slightly as the frequency increases up to 1000 cycles, but may be as low as 30 per cent of the 60-cycle value at radio frequencies, probably on account of the heating effect.

Commercial Oils. For service in which the liquid will be in contact with air, the use of mineral oil has been practically universal because of its stable nature and low cost. Since the breakdown of all oils is approximately 30 to 40 kv rms in a standard 0.1-in. gap between 1-in.-diameter disks, little is to be gained by substituting other oils except a higher dielectric constant, which is obtained at the cost of lower resistivity. For hermetically sealed applications, such as capacitors, purified castor oil, with a constant of 4.7, is often used. Carefully purified, chlorinated hydrocarbons, such as "Pyranol," with a constant of 4.5, are also employed for this purpose and have an effective advantage in being explosion-proof.

Mineral oils must be carefully purified to remove unsaturated compounds which cause accelerated oxidation in service, resulting in low resistivity, low dielectric strength, high power factor, and the rapid development of sludge. Too drastic a purification, however, removes naturally occurring antioxidants in the oil, and stability is decreased. The oxidation of transformer oils may be avoided by the use of oxygen-free atmospheres above the oils as in the "Inertaire" system.

Filtration through diatomaceous earth is effective in increasing the resistivity of many liquids. Filtration through hard papers is commonly used for purifying and drying oils in transformer service. Oil should be filtered when the dielectric strength in the standard 0.1-in. gap drops below 22 kv rms. A good oil will show 30 to 50 kv rms. Low-viscosity oils seem to have the highest dielectric strength, although the flash and fire points usually are lower.

Testing methods for electrical insulating oils have been standardized by the American Society for Testing Materials; see ASTM D117. Typical properties for commercial oils are shown in Table 4.

Table 4. Properties of Commercial Oils

Property	Mineral Transformer Oil	Mineral Capacitor Oil	Castor Oil
Density, average	0.87	0.91	0.96
Viscosity at 37.8° C (100° F), in Saybolt seconds, average	57	100	1400
Flash point, in deg cent, minimum	133	149	....
Fire point, in deg cent, minimum	148	170	....
Pour point, in deg cent, maximum	−40	−40	−15
Neutralization number, in mg KOH per gram, maximum	0.03	0.03	2.0
Dielectric constant	2.2	2.2	4.7
Power factor, at 100° C and 1000 cycles	....	0.0025 max	0.01
Resistivity, ohm-cm, at 100° C	4×10^{11}	$>5 \times 10^{12}$	6.6×10^{10}
Dielectric strength, at 25° C, in kv, minimum	30	30	28
Coefficient of expansion per deg cent	6.3×10^{-4}	6.3×10^{-4}	....
Thermal conductivity, in cal per sec per cm per deg cent	3×10^{-4}	....	4.3×10^{-4}

Synthetic Liquids. Synthetic insulating liquids of the non-inflammable type are known as askarels, and a draft of proposed testing methods has been published in the *Proceedings of the American Society for Testing Materials*, Vol. 43, 353. These liquids consist of mixtures of various chlorinated diphenyls and tri- or dichlorobenzene so adjusted that the pour point is reduced below service temperatures. The dielectric constant is about 4.2. Viscosity at 100 deg fahr is about the same as that of mineral transformer oils. Dielectric strength is slightly higher, and the fact that the liquids are non-inflammable permits the use of large transformers without fireproof vaults. On account of the non-explosive nature of the liquid, the air space above it may be sealed from the atmosphere with a safety diaphragm designed to relieve pressure if a fault occurs. Improved stability to oxidation and sludging is another advantage of these liquids. Trade names: *Pyranol, Inerteen.*

Silicones. Silicone fluids are a recent development and are expensive (as of 1949), but they appear to be most promising as an insulating medium. Silicone liquids consist of chains of alternate silicon and oxygen atoms with various organic groups attached in pairs to the silicon atoms. Those with two methyl groups attached are dimethyl silicones (Dow-Corning DC 200 fluids). The viscosity increases with the chain length. These fluids are suitable for use from −40 deg fahr to 400 deg fahr. Another series (DC 500 fluids) is serviceable from −70 deg fahr to 200 deg fahr. The general advantages of silicone fluids are:

1. Low temperature-viscosity slopes.
2. High flash and fire points.
3. Low volatility and negligible vapor pressure.
4. High resistance to oxidation and heat.
5. Lack of color, odor, or toxicity.
6. Low power factor over a wide frequency range.
7. Non-corrosive to metals and non-solvent for rubber and plastics.

The characteristics of these fluids are shown in Table 5.

Table 5. Properties of Some Liquid Silicones (Dow Corning)

Fluid Type	Viscosity Grade, centistokes at 25° C	Viscosity Temperature Coefficient $\left(1 - \frac{V_{210° F}}{V_{100° F}}\right)$	Freezing Point, deg cent	Boiling Point, deg cent	Flash Point, deg cent min	Specific Gravity 25° C/25° C	Coefficient of Thermal Expansion Unit = 10^{-3}/°C	Refractive Index at 25° C
DC 500	1.0	0.37	−86	$152^{760\text{ mm}}$	37.8	0.818	1.451	1.3822
	3.0	.51	−70	ca. $80^{0.5\text{ mm}}$	107	.896	1.170	1.394
	10.0	.57	−67	$>200^{0.5\text{ mm}}$	176	.940	1.035	1.399
	50.0	.59	−55	$>250^{0.5\text{ mm}}$	282	.955	1.00	1.402
DC 200	100	.60	See Note 1	See Note 2	315	.968	0.969	1.4030
	350	.62	"	"	329	.972	0.966	1.4032
	1000	.62	"	"	337	.973	0.963	1.4035

Note 1. Recommended for use above −40 deg cent.
Note 2. Less than 2 per cent volatile during 48 hours at 200 deg cent.

Electrical Properties of DC 200 Fluids at 25 Deg Cent and 50 Per Cent Relative Humidity

Frequency, cycles per sec	Dielectric Constant	Power Factor
10^3	2.85	0.001
10^6	2.83	.002
10^8	2.81	.006

Dielectric strength, 250–300 volts per mil
Volume resistivity, 1×10^{14} ohm-cm

7. GASES AS DIELECTRICS

Dielectric Constant. The dielectric constants of gases are close to unity and nearly independent of frequency. The change in dielectric constant of dry non-polar gases with temperature or pressure is slight and may be calculated approximately from the equation

$$K - 1 = A \times \frac{p}{273 + t}$$

where A is a constant (2.12×10^{-4} for air).
p is the pressure in millimeters of mercury.
t is the temperature in degrees centigrade.

Table 6. Dielectric Constant of Gases

Gas	Temperature, deg cent	Pressure, atmospheres	Dielectric Constant	Observer
Air	0	1	1.000590	Boltzmann 1875
"	19	20	1.0108	Tangl 1907
"	......	40	1.0218	" "
"	......	60	1.0330	" "
"	......	80	1.0439	" "
"	......	100	1.0548	" "
Argon	23	1	1.000530	Braunmuhl 1927
Carbon dioxide	0	1	1.000985	Klemencic
" "	15	10	1.008	Linde 1895
" "	......	20	1.020	" "
" "	......	40	1.060	" "
Carbon monoxide	0	1	1.000690	Boltzmann
Ethylene	0	1	1.0031	"
Hydrogen	0	1	1.000264	"
Methane	0	1	1.000944	"
Nitrogen	0	1	1.00061	......
Oxygen	0	1	1.00055	......

Conductivity and Ionization. The conductivity of gases at low potential gradients is negligible in the absence of ionizing radiation, such as ultraviolet light or X-rays. If a sufficient voltage gradient exists, all the ions are drawn to the electrodes as fast as they are produced and the very small ion current is constant over a considerable range of gradient. Increasing the gradient beyond this saturation range accelerates the negative ions (or electrons) to a velocity which is sufficient to expel electrons from neutral gas molecules at each collision.

If the number of ions liberated by the collisions exceeds the number of negative ions which are lost by recombination to form neutral molecules and by diffusion out of the field, the collision process is cumulative and the current increases continuously to breakdown. If the electric field is uniform, sparkover will occur, but if the high voltage gradients are confined to a small region, such as the vicinity of a pointed electrode, a local discharge, known as corona, occurs. Local discharges produce visible and ultraviolet light which is effective in increasing ionization throughout the field. This internal photo-ionization causes extremely rapid breakdowns of air gaps at sufficiently large voltage gradients.

Corona. The production of a corona discharge requires a considerable current which is carried by ions of lower velocity in the dark regions of the field. A significant power loss may occur if the corona becomes appreciable. Corona is objectionable because it causes radio-frequency interference and produces ozone if oxygen is present. Ozone and the ultraviolet light from the discharge cause rapid deterioration of many solid dielectric materials, especially of rubber. Corona is normally prevented by operation at reduced voltages or by the use of suitable corona shields which are designed to produce a more nearly uniform voltage gradient throughout the air space.

Dielectric Strength and Sparkover. Dielectric strength is the maximum potential gradient at the instant of sparkover or at the onset of corona. The gradient is determined by the geometry and spacing of the electrodes. Dielectric strength is influenced by the nature and purity of the gas, by the density of the gas, and to a lesser degree by the electrode material. When the mean free ion path between collisions is lengthened by lowering the gas density, the critical terminal velocity necessary for ionization by collision is attained with a lower potential gradient. The reduction in dielectric strength with decreasing density continues until the number of atoms between the electrodes is so small that very few collisions occur. If the density is still further reduced and the electric field is non-uniform, the sparkover will occur over some path longer than the shortest distance between electrodes. If the field is uniform, the sparkover voltage will increase as the density decreases to very low values, until, in a high vacuum, gradients as high as 6000 kv per cm may be obtained. The minimum in the sparkover voltage is independent of electrode spacing for uniform fields and depends solely on the nature and purity of the gas and on the electrode material. A typical set of values for various gases is shown in Table 7.

Table 7. Minimum Sparking Potentials

Gas	Volts (d-c)	Gas	Volts (d-c)
Air	341	Hydrogen	278
Carbon dioxide	419	Nitrogen	251
Helium	261	Oxygen	455

Sparkover voltage in uniform fields is a nearly linear function of the product of gas density and electrode spacing (Paschen's law).

An empirical equation which reproduces the entire sparkover curve in air for uniform fields, including the region around the minimum sparking potential, is

$$V = \frac{404pS \dfrac{293}{273 + t}}{3.000 + \log_e \left(pS \dfrac{293}{273 + t} \right)} + 300 \text{ volts (crest)}$$

where p is the pressure in millimeters of mercury.
S is the electrode spacing in centimeters.
t is the temperature in degrees centigrade.

The calculation of the sparkover voltage for non-uniform fields is not simple; see F. W. Peek, *Dielectric Phenomena in High Voltage Engineering*, McGraw-Hill (1929) for extensive formulas and tables for this purpose. At spacings of the order of one-half the sphere radius, the sparkover voltage of a gap between equal spheres is slightly higher than for a uniform field, but as the ratio of spacing to radius increases the sparkover voltage becomes much less than for a uniform field. At ratios above 2, corona occurs before sparkover, but corona is not easily detected at 60 cycles until the ratio is about 8. Representative sparkover voltages for sphere gaps are shown in Table 8.

Table 8. Sparkover Voltage in Rms Kilovolts

Barometer 76 cm, temperature 25 deg cent

Diameter of Spheres

Gap Spacing		2 cm		6.25 cm		12.5 cm		25 cm	
cm	in.	NG *	G †	NG	G	NG	G	NG	G
0.2	0.079	5.6	5.6	...	...	...	...	...	...
0.25	.098	...	...	...	...	6.5	6.5	...	...
0.3	.118	8.0	8.0	...	...	...	...	...	...
0.4	.158	10.3	10.3	...	...	...	...	...	...
0.5	.197	12.5	12.5	12.0	12.0	12.0	12.0	11	11
0.6	.236	14.8	14.6	...	...	...	...	...	...
0.7	.276	17.0	16.7	...	...	...	...	...	...
0.8	.315	18.9	18.6	...	...	...	...	...	...
0.9	.354	20.8	20.2	...	...	...	...	...	...
1.0	.394	22.6	21.7	22.5	22.5	22.0	22.0	22	22
1.2	.472	25.9	24.4	...	...	...	...	...	...
1.4	.551	28.9	26.4	...	...	...	...	...	...
1.5	.591	...	...	31.5	31.5	31.5	31.5	32	32
1.6	.630	...	28.2	...	...	...	...	...	...
1.8	.709	...	30.0	...	...	...	...	...	...
2.0	.787	...	...	41.0	41.0	41.0	41.0	42	42
2.5	.984	...	...	...	...	...	...	52	52
3.0	1.181	...	...	57.5	56.0	59.0	59.0	61	61
4.0	1.575	...	...	70.5	66.0	76.0	75.0	78	78
5.0	1.969	...	...	81.0	73.0	91.0	89.0	96	94
6.0	2.362	...	...	89.0	79.0	105.0	102.0	112	110
7.0	2.756	...	...	96.0	83.0	118.0	112.0	...	...
8.0	3.15	...	...	102.0	88.0	130.0	120.0	...	...
9.0	3.543	...	...	107.0	90.5	141.0	128.0	...	...
10.0	3.937	...	...	110.0	93.0	151.0	135.0	171	166
12.0	4.724	...	...	...	...	167.0	147.0	...	...
12.5	4.921	...	...	...	...	...	...	203	196
15	5.906	...	...	...	...	188.0	160.0	230	220
17.5	6.89	...	...	...	...	201.0	168.0	255	238
20	7.874	...	...	...	...	213.0	174.0	278	254
22.5	8.858	...	...	...	...	...	...	297	268
25	9.843	...	...	...	...	...	...	314	280
30	11.811	...	...	...	...	...	...	339	300
40	15.748	...	...	...	...	...	...	385	325

Values from F. W. Peek, *Dielectric Phenomena in High Voltage Engineering*, 3d ed., McGraw-Hill Book Co. (1929).

* NG = electrodes balanced to ground.
† G = one electrode grounded.

Table 9 shows the influence of altitude on sparkover voltage at constant temperature for uniform fields in gaps of various lengths. The table also is approximately correct for closely spaced sphere gaps. Correction for the lower air temperatures shown is seldom warranted unless the actual temperature of the air in the gap is known.

Table 9. Sparkover Voltages at High Altitudes

Ratio of sparkover voltage to sea-level sparkover in uniform fields

Altitude, Unit = 1000 ft	Standard Air Temp, 0° C	Pressure		Relative Pressure	Sparkover Voltage at Constant Temperature			
					Gap Spacing, cm			
		mm Hg	in. Hg		0.01	0.1	1.0	10.0
0	+15.0	760.0	29.92	1.000	1.000	1.000	1.000	1.000
5	+5.1	632.2	24.89	0.832	0.910	0.862	0.85	0.845
10	−4.8	522.6	20.58	0.687	0.828	0.742	0.719	0.712
15	−14.7	428.8	16.88	0.565	0.759	0.637	0.603	0.593
20	−24.6	349.1	13.75	0.459	0.695	0.545	0.505	0.493
25	−34.5	281.9	11.10	0.371	0.640	0.468	0.418	0.390
30	−44.4	225.6	8.88	0.296	0.592	0.399	0.345	0.331
35	−54.3	178.7	7.04	0.247	0.552	0.340	0.283	0.268
40	−55.0	140.7	5.54	0.185	0.517	0.291	0.232	0.217
45	−55.0	110.8	4.36	0.146	0.488	0.251	0.190	0.175
50	−55.0	87.3	3.44	0.115	0.465	0.219	0.159	0.141

The dielectric strength of gases is subject to large changes with impurities. At low pressures, 0.1 per cent of argon in neon gas reduces the dielectric strength by 75 per cent. Mercury vapor likewise lowers the sparkover voltage. At atmospheric pressure, the addition of small amounts of carbon tetrachloride or chloroform vapor increases the dielectric strength by 50 per cent. Table 10 shows the approximate relative dielectric strength of gases. Although Freon (dichlorodifluoromethane) has a higher relative dielectric strength, its vapor pressure varies from 9.3 psi at −40 deg cent to 139 at +40 deg cent, whereas the pressure with sealed-in nitrogen increases by only 35 per cent in this interval. Furthermore, if the Freon does break down, its decomposition products are corrosive and carbon is deposited. Dry, oil-pumped nitrogen is undoubtedly the best gas dielectric unless cooling is involved, which may necessitate the use of hydrogen.

Table 10. Approximate Relative Dielectric Strength of Gases

Gas	Pressure in Absolute Atmospheres				
	1/3	1	2	4	8
Nitrogen	1.0	1.0	1.0	1.0	1.0
Air	0.9	0.9	0.9	0.9	1.0
Oxygen	0.9	0.8	0.8	0.8	....
Carbon dioxide	0.9	0.9	0.9	0.9	0.8
Hydrogen	0.6	0.6	0.6	0.6	....
Freon	....	2.4	2.2	2.2	2.2
Nitrogen plus carbon tetrachloride vapor	....	1.6	1.5	1.3	1.2
Helium	0.3	....	....	....	....

At frequencies above 10 kc the dielectric strength of air decreases slightly. Owing to the low velocity of positive ions, they are not swept to the electrodes during one half cycle, so that they remain to distort the field on the next half cycle. At frequencies above 60 kc, a reduction of 7 to 13 per cent in breakdown voltage is to be expected, but no further decrease occurs up to at least several megacycles. If the gap is illuminated with ultraviolet light, a decrease of 17 to 20 per cent is obtained.

MAGNETIC MATERIALS

By R. M. Bozorth and R. A. Chegwidden

Of all the common elements only iron, cobalt, and nickel have magnetic properties greatly different from those of air or vacuum. The magnetic materials in common use consequently contain at least one of these elements and sometimes all three. These materials, called *ferromagnetic*, make possible the operation of motors, generators, transformers, and practically all electromagnetic devices.

Non-ferromagnetic materials are of little importance on account of their magnetic properties. They fall into two classes: *paramagnetic* materials which are but slightly more magnetic than a vacuum, and are therefore attracted weakly by the poles of an electromagnet, and *diamagnetic* materials which are repelled weakly by an electromagnet Because the magnetic materials used in communication are almost without exception ferromagnetic materials, this article will be devoted to a description of this class. They may be divided into "high permeability" (or magnetically "soft") and "permanent magnet" (or magnetically "hard") materials. Ferromagnetic materials are commonly referred to as simply "magnetic" materials.

The magnetic properties of materials depend primarily on the nature of the atoms which compose them. In magnetic materials the atoms are small permanent magnets that owe their magnetic moments to uncompensated spinning electrons lying in electron shells inside the atom. For a material to be ferromagnetic these shells must be incomplete (i.e., have spaces for more electrons than are present) and have an excess of electrons spinning in one direction, and the atoms must be arranged in regular fashion on a space lattice, with atom centers not too close together. The important elements whose atoms fulfil these conditions occur in one part of the periodic table—they are iron, cobalt, and nickel. Gadolinium has been found to be ferromagnetic, and manganese and chromium can also give rise to ferromagnetism when alloyed or chemically combined with the right non-ferromagnetic elements. The *Heusler alloys* are composed of manganese, aluminum, and copper; in comparison with many of the alloys of iron, cobalt, and nickel, they are quite inferior from a practical magnetic standpoint and have had no commercial use.

Magnetic properties depend also on crystal structure, state of strain, temperature, and other factors.

8. MAGNETIC CHARACTERISTICS

MAGNETIZATION AND PERMEABILITY CURVES. The properties of magnetic materials are usually described first of all by a *magnetization curve* such as that shown in Fig. 1. Here the *magnetic induction*, B, in a ring sample is plotted against the *magnetizing force*, H.

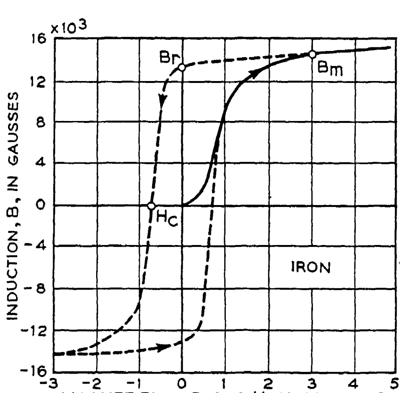


FIG. 1. Magnetization Curve and Hysteresis Loop for Annealed Iron Ring

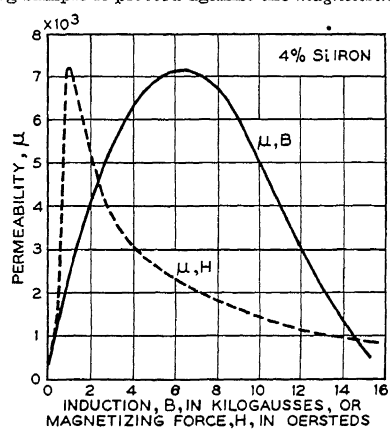


FIG. 2. Typical Permeability Curves for Hot-rolled 4 Per Cent Silicon Iron

B is a measure of the amount of the magnetization; it is defined specifically below under "Definitions." H represents the magnetizing force required to produce the magnetic induction, B; it is usually measured in oersteds or in ampere-turns per inch. The magnetic induction is sometimes described in terms of the intensity of magnetization, I, equal to $(B - H)/4\pi$. The increase in induction due to the material alone is $B - H$, sometimes called the *intrinsic induction;* this quantity becomes important when the magnetizing force is high, e.g., when determining B_s, the saturation induction, highest attainable value of $B - H$ in a material.

The ease with which a magnetic material can be magnetized is measured by the ratio B/H, called the *permeability*, μ. Typical permeability curves plotted against B and H are shown for a sample of 4 per cent silicon-iron in Fig. 2.

In communication work, the permeability in low fields is especially important. Figure 3 shows the characteristic μ, H curves for several common materials; in the lowest fields the curves usually become straight lines.

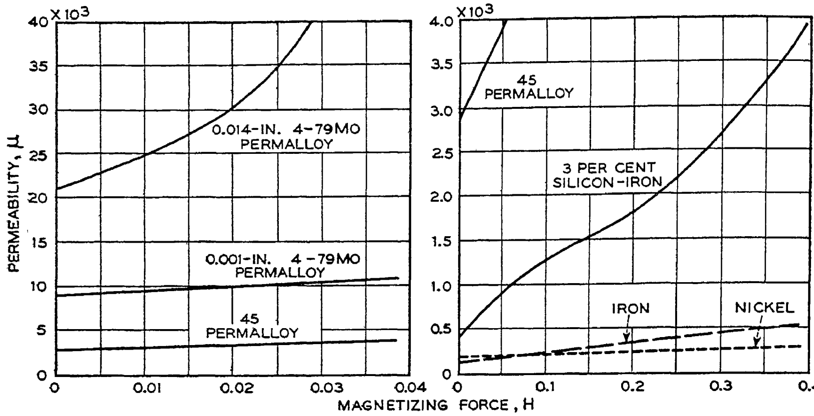


FIG. 3. Characteristic μ vs H Curves at Low Magnetizing Forces

Hysteresis Loops. Other important properties of magnetic materials are shown by the hysteresis loop produced when B is plotted against H, as H is increased to a maximum, decreased to zero, increased to a maximum in the negative direction, again reduced to zero, and finally increased to the first maximum as shown by the dotted line in Fig. 1. The values of B and of H at which this curve crosses the axes are called respectively the residual induction, B_r, and coercive force, H_c, as indicated in the figure. The area enclosed within the hysteresis loop is a measure of the magnetic energy transferred into heat during the cycle and is designated W_h. Quantitatively

$$W_h = \frac{1}{4\pi} \oint H\, dB$$

in ergs per cm³ when B and H are in gausses and oersteds, respectively.

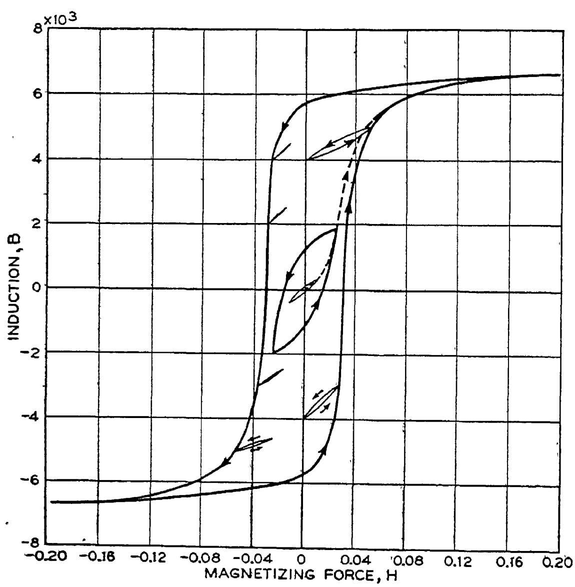


FIG. 4. Minor Hysteresis Loops Shown at Various Points on the Major Loop for a Specimen of 4–79 Molybdenum Permalloy

Demagnetization Curve. That part of the hysteresis loop that lies in the second quadrant, extending from B_r to H_c in Fig. 1, is called the demagnetization curve and is especially important for the description of permanent-magnet materials. This is described more fully in Article 10 and Fig. 10. Hysteresis loops which do not have equal excursions of H (and B) in opposite directions are unsymmetric or minor loops. Some of these are shown in Fig. 4. In a minor loop the ratio of the total change in B, ΔB, to the total change in H, ΔH, is called the incremental permeability $\mu_\Delta = \Delta B/\Delta H$. In the limit as ΔB and ΔH approach 0, the incremental permeability is the *reversible permeability*, μ_r. This is sometimes referred to as superposed permeability, and is discussed in more detail in Article 15 (Fig. 24).

Air Gaps. When an air gap is cut in a closed magnetic circuit, such as a ring sample, magnetic poles are produced on either side of the break, and the apparent permeability of the material is reduced because of the reluctance of the gap. The effect of even a small air gap in a high-permeability circuit may be very appreciable. See Fig. 5, and Fig. 19 of Article 15.

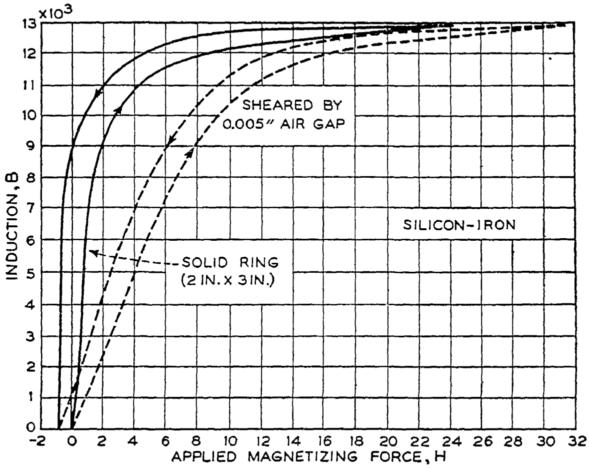


FIG. 5. Magnetization Curves and Parts of Hysteresis Loops Showing Effect of 0.005-in. Air Gap Cut in a Silicon Iron Ring 2-in. I.D. by 3-in. O.D.

Eddy-current Loss. When magnetic fields are varied with some rapidity, as they are in most machinery, the material is subject not only to the hysteresis loss already described but also to eddy-current loss. This results from the flow of electric currents within the material, induced by the changing flux in them. They increase with increase in the frequency, conductivity, and permeability of the material. Since time is required for these currents to build up and to decay, the application of a varying field is accompanied by a delay in the corresponding magnetic induction. The summation of the hysteresis and eddy-current losses is frequently called the core loss or iron loss of the material.

DEFINITIONS. The following definitions are taken from ASTM Specification A127 which may be referred to for definitions of other terms relating to magnetic materials. This list is in alphabetical order, for reference only.

Coercive Force. H_c. The magnetizing force required to bring the induction to zero in a magnetic material which is in a symmetrically cyclically magnetized condition. The *coercivity* is that property of a material measured by the maximum value of the coercive force.

Induction, Intrinsic. B_i. The excess of the induction in a magnetic material over the induction in vacuum, for a given value of the magnetizing force. The equation for intrinsic induction is

$$B_i = B - H$$

Induction, Magnetic (Magnetic Flux Density). B. Flux per unit area through an element of area at right angles to the direction of the flux. The cgs unit of induction is called the gauss (plural gausses) and is defined by the equation:

$$B = \frac{d\phi}{dA}$$

Under a-c conditions B_{max} may be calculated as follows:

$$B_{max} = \frac{E \times 10^8}{4\,ff\,NAf}$$

where E is in rms volts; ff is the form factor.

Induction, Normal. B. The limiting induction, either positive or negative, in a magnetic material which is in a symmetrically cyclically magnetized condition.

Induction, Residual. B_r. The magnetic induction corresponding to zero magnetizing force in a magnetic material which is in a symmetrically cyclically magnetized condition. The *retentivity* is the property of a magnetic material measured by the maximum value of the residual induction.

Induction, Saturation. B_s. The maximum intrinsic induction possible in a material.

Magnetic Flux. ϕ. A condition in a medium produced by a magnetomotive force, such that when altered in magnitude a voltage is induced in an electric circuit linked with the flux. The cgs unit of magnetic flux is called the maxwell and is defined by the equation:

$$e = -N\frac{d\phi}{dt} \times 10^{-8}$$

where e = induced emf in volts, and $d\phi/dt$ = time rate of change of flux in maxwells per second.

Magnetizing Force. H. Magnetomotive force per unit length. The cgs unit is called the oersted and is defined by the equation:

$$H = \frac{dF}{dl}$$

where F is in gilberts and l in centimeters. For a toroid, or at the center of a long solenoid, the magnetizing force in oersteds may be calculated as follows:

$$H = \frac{0.4\pi NI}{l}$$

where I is in amperes and l is in centimeters.

Magnetomotive Force. F. That which tends to produce a magnetic field. In magnetic testing it is most commonly produced by a current flowing through a coil of wire, and its magnitude is proportional to the current and to the number of turns. The cgs unit of magnetomotive force is called the gilbert and is defined by the equation:

$$F = 0.4\pi NI$$

where I is in amperes. Magnetomotive force may also result from a magnetized body.

Permeability, a-c. μ_{ac}. A-c permeability is variously defined, and the values obtained for a given material depend on the methods and conditions of measurement. As measured by the Standard Methods of Test for Magnetic Properties of Iron and Steel (ASTM Designation A34), it is the ratio of the maximum value of induction to the maximum value of the magnetizing force for a material in a symmetrically cyclically magnetized condition.

It is sometimes defined as the ratio of the rms flux density to the rms magnetizing force. Some of the factors which affect a-c permeability are thickness of laminations, frequency, and resistivity.

Permeability, Incremental. μ_Δ. The ratio of the cyclic change in magnetic induction to the corresponding cyclic change in magnetizing force when the mean induction differs from zero.

Permeability, Initial. μ_0. The slope of the normal induction curve at zero magnetizing force.

Permeability, Normal. μ. The ratio of the normal induction to the corresponding magnetizing force. In the cgs system the flux density in a vacuum is numerically equal to the magnetizing force, and, consequently, the magnetic permeability is numerically equal to the ratio of the flux density to the magnetizing force. Thus:

$$\mu = \frac{B}{H}$$

Note: In a non-isotropic medium the permeability is a function of the orientation of the medium, since, in general, the magnetizing force and the magnetic flux are not parallel.

Permeability, Reversible. μ_r. The incremental permeability when the cyclic change in induction is vanishingly small.

9. HIGH-PERMEABILITY MATERIALS

Preparation and Heat Treatment. In many applications such as motors, generators, transformers and relays, it is desirable to have materials of high permeability. The common materials used are iron, silicon-iron alloys, and iron-nickel alloys to which various other metals have been added. Commercial materials are usually melted in open-hearth or electric furnaces, poured to form ingots, then rolled to slabs and finally to sheets or rods of the required dimensions. After fabrication into the final form in which they are to be used, they must be subjected to a heat treatment which is appropriate to the particular alloy, in order to develop their best magnetic qualities. These heat treatments usually consist of heating to some temperature lying between 800 and 1200 deg cent and cooling at a definite rate to room temperature. Figure 6 shows magnetization curves for ingot iron, (1) as hot and cold rolled to final size, (2) as annealed at 900 deg cent for 1 hour, and (3) after heat treatment at the unusually high temperature of 1400 deg cent for 6 hours. The anneal at 900 deg cent may be referred to as a strain-relief anneal; that at the higher temperature as a purifying anneal because during the process some of the impurities have been removed from the iron.

In processes involving annealing, account must be taken of phase transformations and "order-disorder" phenomena (orderly arrangements of the atoms) like those found in iron and Permalloy, respectively. In some materials, such as the nickel-manganese alloys, a magnetic material may be made non-magnetic by cooling rapidly from a high temperature and thus preserving the disorderly distribution of nickel and manganese atoms stable at this temperature. In the iron-nickel alloys containing 50 to 80 per cent nickel, marked changes in properties are produced by annealing in the presence of a magnetic field.

Magnetic Elements. Of the magnetic elements, iron, cobalt, and nickel, iron is the only one used commercially to any considerable extent in unalloyed form. It is made in large tonnages for motors, generators, and relays. Its characteristic properties are described in Fig. 9 and Table 1. Nickel finds a limited use because of its magnetostrictive properties, for example in supersonic underwater apparatus.

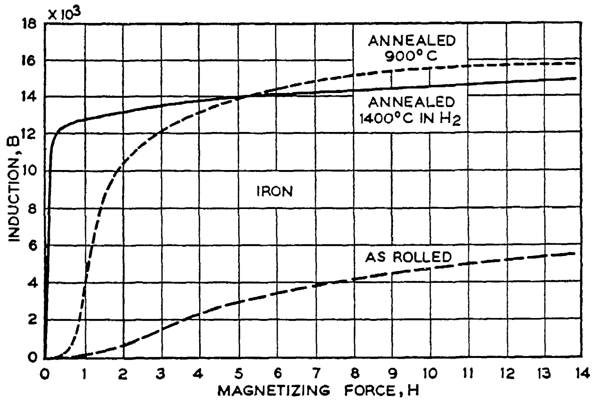


FIG. 6. Effect of Heat Treatment on the Magnetic Properties of Iron

Iron-silicon Alloys. Next to unalloyed iron, these alloys are used in the greatest quantities, in power transformers, motors, generators, relays, and receivers. The addition of silicon increases the resistivity and so cuts down the power loss due to eddy currents, and it also has some effect in increasing the permeability and decreasing the hysteresis loss. Various commercial grades are available containing up to about 5 per cent silicon. Great advances have been made in the last few years by fabricating the sheet by cold rolling instead of hot rolling. As with most magnetic materials, the iron-silicon crystals are most easily magnetized along one particular crystallographic direction, and the cold rolling and associated heat-treating processes are adjusted to orient the crystals so that as many as possible have their directions of easy magnetization aligned in the direction of rolling. Material made in this manner is called *grain oriented.* The grain-orienting process

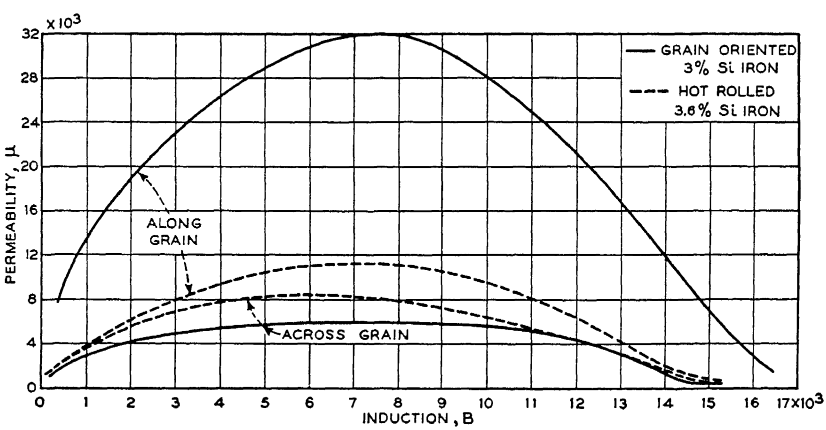


FIG. 7. Comparison of the μ *vs* B Curves for Samples of Hot-rolled and Grain-oriented Silicon Iron

increases the permeability and reduces hysteresis loss. The increase in permeability, however, is effective only when the sheet is magnetized in the direction in which it is rolled, and in most other directions the permeability is lowered. Grain-oriented silicon-iron is available in grades containing up to slightly over 3 per cent silicon. Properties of the grain-oriented material containing 3 per cent silicon are compared with the hot-rolled product in Fig. 7.

Iron-nickel Alloys. These alloys are used when particularly high quality is desired, usually in transformers of various kinds and in magnetic shields. Their permeabilities are much greater than those of other alloys, and they have high resistivity and low energy loss. The most important binary alloys contain 78 per cent nickel and 45 to 50 per cent nickel. The former, called *78 Permalloy* or Permalloy A, has a high initial permeability (about 10,000) and a low coercivity (0.05) and requires special heat treatment for the development of these properties. After the usual anneal at 1000 to 1100 deg cent it must be cooled rapidly in order to develop maximum quality. The iron-nickel alloys containing 45 to 50 per cent nickel are useful in certain relays, transformers, receivers, and other apparatus. They have reasonably high permeabilities and incremental permeabilities, and their saturation inductions and resistivities are higher than those of the 78 per cent alloy. The cooling rate from the annealing temperature is not critical for these alloys.

Another development in this field is the grain-oriented 50 per cent nickel-iron alloy, originated in Germany and called Permenorm 5000-Z. After special heat treatment, this material exhibits hysteresis loops which are practically rectangular. The grain orientation is accomplished by a 99 per cent cold reduction before the final heat treatment.

When higher resistivities are required, other elements are added to the iron-nickel alloys, and it is often found that the resulting alloy has also higher initial and maximum permeabilities. One of the most useful of these is *Molybdenum Permalloy* containing 4 per cent molybdenum and 79 per cent nickel. Similar properties are obtained in *Mumetal* containing 2 per cent chromium, 5 per cent copper, and 75 per cent nickel. Initial permeabilities of 20,000 to 30,000 and maximum permeabilities of about 100,000 are often found in these alloys. The molybdenum Permalloy must be heat-treated under non-oxidizing conditions and preferably cooled at a definite rate; the Mumetal must be heat-treated in a hydrogen atmosphere for best results. An alloy containing 5 per cent molybdenum and 79 per cent nickel is called *Supermalloy* and has an initial permeability of 50,000 to 150,000 and a maximum permeability of about 1,000,000. These properties are obtained by controlled melting of suitable raw materials, heat treating in hydrogen at 1300 deg cent, and cooling at a critical rate.

Some interesting properties have been obtained in alloys containing nickel, iron, and cobalt, called *Perminvars*. A typical alloy contains 45 per cent nickel, 25 per cent cobalt, and 30 per cent iron, and its permeability is characteristically independent of magnetizing force over a relatively large range. To retain this property, however, it must never be magnetized above this range. Above the range of constant permeability the hysteresis loops have peculiar constricted forms with very low residual induction at intermediate field strengths. At present these alloys are not being used commercially.

Iron-cobalt Alloys. Alloys of iron with approximately 35 per cent to 55 per cent cobalt are remarkable in that the saturation, B_s, is higher than that of either iron or cobalt. One of these alloys, containing equal parts of iron and cobalt, is called *Permendur;* it is useful where high flux densities are necessary as in the pole tips of electromagnets. The binary alloys of iron and cobalt can be hot rolled and machined, but they are very difficult to cold roll. *Vanadium Permendur* contains 2 per cent vanadium and equal parts of iron and cobalt; the addition of vanadium makes it possible to cold-roll the alloy to thin sheets. In this form it has found an important application in telephone diaphragms. Vanadium Permendur can be machined and even punched after cold rolling, but it becomes somewhat brittle when annealed. At high inductions the superposed permeability of the iron-cobalt alloys is higher than that of any other material (see Fig. 24).

Other Alloys. Among the other alloys that have high permeability may be mentioned aluminum-iron, molybdenum-iron, and *Sendust*, the last developed by the Japanese. *Aluminum-iron* containing up to about 6 per cent aluminum has been produced having magnetic properties somewhat better than the iron-silicon alloys, but because of manufacturing difficulties it has never been popular. In Japan high permeabilities have been obtained in the alloy containing 14 to 15 per cent aluminum, called *Alfer*. *Iron-molybdenum* alloys also have good magnetic permeabilities at low and moderate flux densities and are more ductile than the iron-silicon alloys. Although they are not in common use, at least one manufacturer is planning to market them. *Sendust* contains about 9 per cent silicon, 5 per cent aluminum, and the rest iron. Both initial and maximum permeability are high, the initial sometimes as high as 35,000. *Sendust* is an unusual high-permeability material in that it is quite brittle and must be cast and ground to finished size. Because of its brittleness, *Sendust* has been formed into powder and used, especially in Japan, in pressed powder cores for loading coils and other coils for high-frequency circuits.

Other materials are useful in powdered form for applications of this kind. Most important among them are *powdered Permalloy* and *Carbonyl iron*. The former has the nominal composition of 2 per cent molybdenum, 80 per cent nickel, and 18 per cent iron; it is melted without the addition of a deoxidizer or desulfurizer, and after hot rolling it can

Table 1. Characteristic Properties of Various High-permeability Materials

Material	Approximate Composition, per cent	Heat Treatment Temperature	Permeability at $B = 20$ Gausses, μ_{20}	Maximum Permeability, μmax	Saturation, B_s gausses	Hysteresis Loss at Saturation, W_h ergs/cm^3	Coercivity, H_c oersteds	Resistivity, ρ microhm-cm	Density, δ gm/cm^3
Magnetic iron	99.9Fe	950° C	200	5,000	21,500	5,000	1 0	10	7.88
4% Silicon iron (hot rolled)	4Si 96Fe	800° C	400	7,000	19,700	3,500	0.5	60	7.6
3% Silicon iron (grain oriented)	3Si 97Fe	800° C	1,500	30,000	20,000	...	0.15	47	7.65
78.5 Permalloy *	78.5Ni 21.5Fe	1,050° C + 600° C	8,000	100,000	10,700	200	0.05	16	8.6
45 Permalloy *	45Ni 55Fe	1,050° C	2,500	25,000	16,000	1,200	0.3	45	8.17
47 to 50% Ni-Fe †	49Ni 51Fe	1,200° C, H_2	4,000	70,000	16,000	220	0.04	35	8.25
4-79Mo Permalloy *	4Mo 79Ni 17Fe	1,100° C	20,000	80,000	8,700	200	0.04	55	8.72
Mumetal *	75Ni 2Cr 5Cu 18Fe	1,200° C, H_2	20,000	130,000	6,500	...	0.05	62	8.58
Supermalloy *	5Mo 79Ni 16Fe	1,300° C, H_2	100,000	800,000	8,000	...	0.004	60	8.77
Sendust *	9Si 5Al 85Fe	Cast	30,000	120,000	10,000	100	0.05	80	7.1
Permendur *	50Co 50Fe	925° C	800	5,000	24,500	12,000	2.0	7	8.3
Hiperco *	35Co 64Fe ‡	850° C	1,000	10,000	24,200	...	1.0	25	8.0
V Permendur *	49Co 49Fe 2V	800° C	800	4,500	24,000	6,000	2.0	26	8.2
2-81Mo Permalloy * (powder)	81Ni 2Mo 17Fe	650° C	125	130	...	...	...	10^6	...
Carbonyl * iron (powder)	99.9Fe		55	132	...	...	...	...	...
Ferroxcube * III	$MnZnFe_4O_8$		1,000	1,500	2,500		1.0	10^8	5.0
Lomu *	6–10Ni 8–12Mn <1Cr 78–85Fe	800° C	1.06	1.06	...	...	...	74	8.03

* Trade names.
† Trade names are: "4750 Alloy," "Hipernik," "Carpenter 49," "Armco 48," etc.
‡ +1% deoxidizer.

be crushed to a fine powder. It is then mixed with a small amount of insulation and pressed into a solid core and heat-treated. The permeability varies, depending upon the amount of insulation, from about 10 to 125. Both hysteresis and eddy-current losses are extremely small, and the permeability is constant over a large range of magnetizing force.

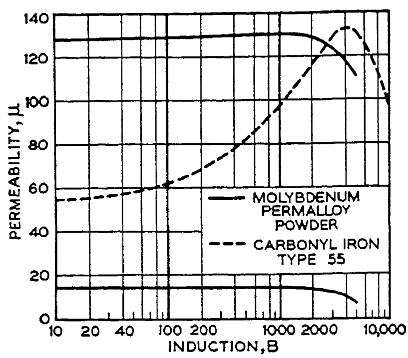


Fig. 8. Examples of Permeability Curves for Cores of Molybdenum Permalloy and Iron Carbonyl Powders

Carbonyl iron powder, prepared by decomposing iron carbonyl, $Fe(CO)_5$, in the vapor phase, consists of small spheroids which are later insulated and pressed. All these powdered materials are used in the cores of coils operating at high frequency; Fig. 8 shows some of their properties.

Often, when the frequency is greater than 1 megacycle, a powder is used composed of one of the iron oxides Fe_2O_3 or Fe_3O_4. Also, properly treated mixtures of complex ferrites, e.g., manganese and zinc ferrites, have been found to have very high electrical resistivities and initial permeabilities of the order of 1000. Materials of this type, developed in Holland under the name *Ferroxcube*, have been used advantageously for certain types of high-frequency coils. The ferrites in general have much higher initial permeabilities but less stability with temperature as compared with the insulated powder materials.

Many of the properties of the most important commercial magnetic materials of high permeability are collected in Table 1 and Fig. 9.

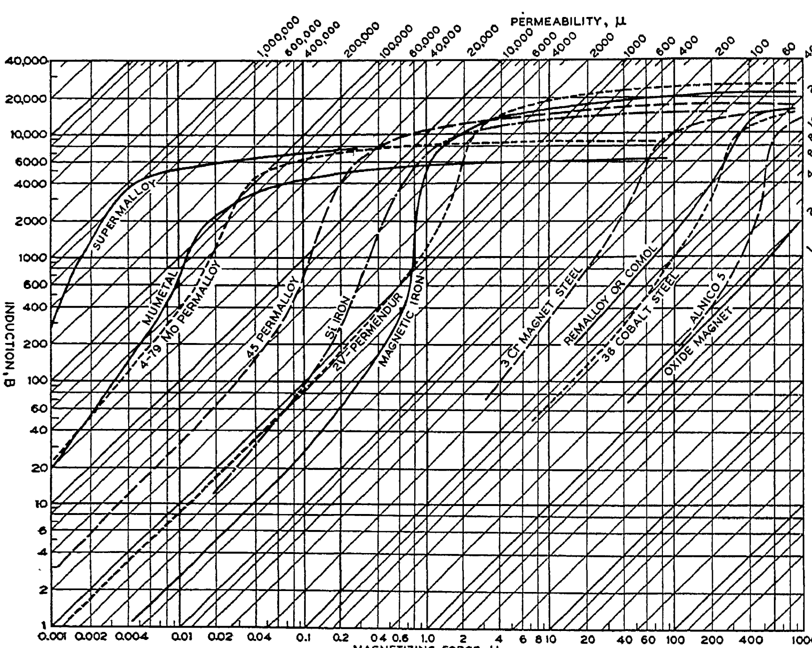


Fig. 9. Magnetization and Permeability Curves for Magnetic Materials in Common Use

Non-magnetic Materials. It is often desirable in dealing with magnetic circuits to have a steel for structural purposes which is non-magnetic. One commonly used material of this sort, called *Lomu*, contains about 10 per cent manganese, 8 per cent nickel, and the remainder iron; chromium and silicon are also sometimes added. Another material is of the stainless-steel type containing 20 to 25 per cent nickel and 25 per cent chromium.

10. PERMANENT-MAGNET MATERIALS

Permanent magnets are useful for their ability to maintain a magnetic field in space without the aid of an external source of power. Since there is no heating, and since modern alloys can produce very high fields, they are used extensively in loudspeakers, small generators, etc. The properties of these materials are best described by means of the demagnetization curve already mentioned. Typical curves for many of the common materials are shown in Fig. 10. In evaluating materials it is also desirable to use an *energy*

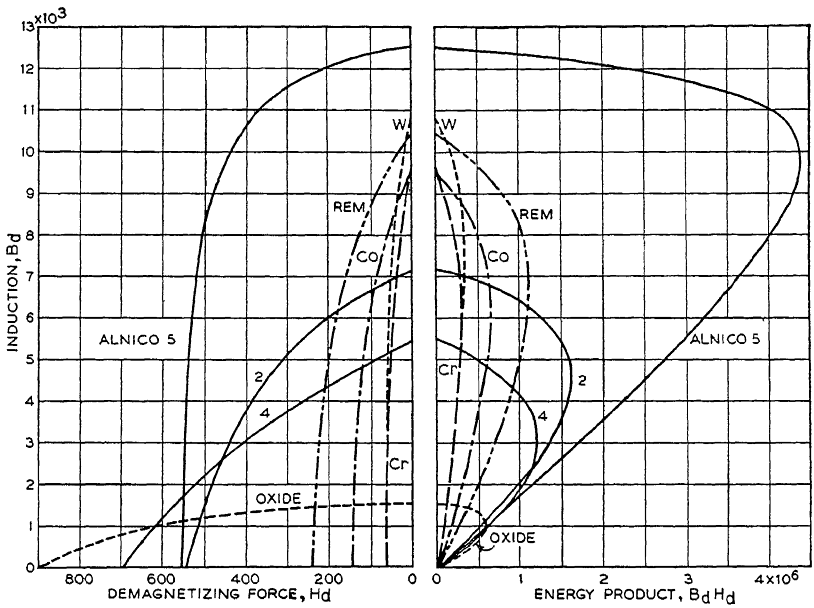


FIG. 10. Demagnetization and Energy Product Curves of Some Important Permanent-magnet Materials. The materials corresponding to the abbreviations may be recognized by reference to Table 2.

product curve formed by plotting the product of B and H, BH, against the magnetic induction B, for points on the demagnetization curve; this is plotted also in Fig. 10. Such a curve is a measure of the energy that can be stored in the magnet, and the best single criterion for the value of a permanent-magnet material is the maximum value of this product, designated $(BH)_m$. In the design of magnetic circuits involving permanent magnets an attempt is usually made to have the magnetic induction in the magnet equal to the B for which (BH) is equal to $(BH)_m$.

The magnetic behavior of a magnet may be described by reference to Fig. 11. The line OA depends on the dimensions of the magnet and is called the load line; it is fixed by the demagnetizing action of the air gap in the magnetic circuit. When the external field has been removed, B and H will be determined by some point on this line, preferably the point for which (BH) is a maximum. It is common practice to "stabilize" a magnet by applying a small negative magnetizing field (point C) and then removing it (point D). Extraneous disturbing fields will then cause changes in induction corresponding to minor loops such as $CDEC$. The minor loops in the third quadrant, under the demagnetizing curve, have slopes approximately the same as the slope of the demagnetizing curve just below the point B; they are important in predicting the changes in induction that occur, e.g., in generators.

Permanent-magnet materials may be classified under the following headings:

- Carbon steel with or without alloying elements.
- Dispersion-hardened alloys.
- Types heat-treated in a magnetic field.
- Ductile alloys.
- Powdered materials.
- Miscellaneous special materials

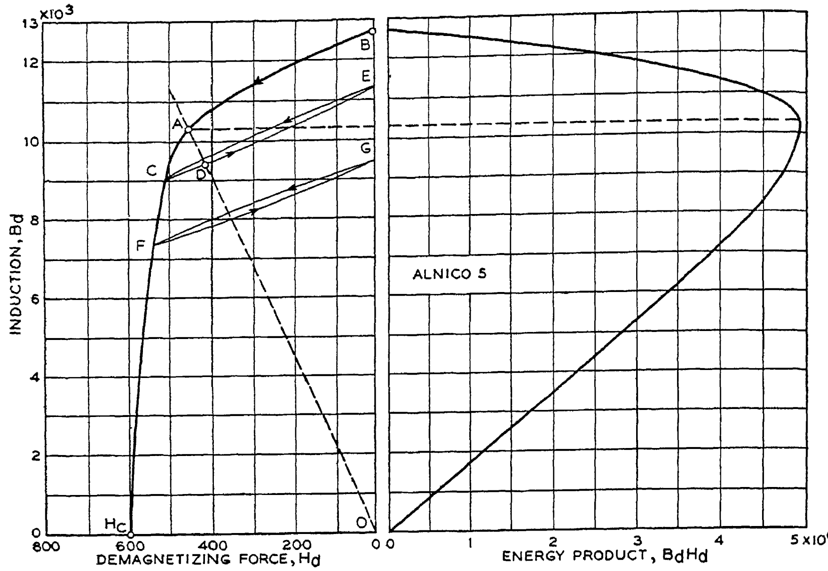


FIG. 11. Curves Useful in the Design of Permanent Magnets. See text.

Carbon Steels. The carbon steels are so called because they depend upon carbon compounds for their permanent-magnet qualities. These materials are usually prepared by hot rolling to finished size and quenching from about 800 to 950 deg cent in water or oil. Their permanent-magnet characteristics tend to deteriorate with time, and it is customary to pre-age such magnets before use by maintaining them for many hours at temperatures between 100 and 150 deg cent. Although their magnetic properties are not of very high quality, many of the carbon-steel magnet alloys are very useful because of their low cost.

The most common steels are carbon-manganese, chromium, tungsten, and cobalt steels. There is a great variety of materials of this kind containing various percentages of the alloying elements and about 0.8 per cent carbon. Representative alloys have been included in Table 2. The coercive force and energy product, as well as the cost, usually increase with the alloying element. The highest energy product obtained in this group is about 1.0×10^6.

Dispersion-hardening Alloys. These alloys contain no essential carbon but depend for their hardening upon the precipitation of one solid phase in another. They are ordinarily heated to 1300 deg cent and quenched in air or oil and are subsequently maintained at 600 to 700 deg cent for several hours; they are therefore often referred to as "age-hardening" alloys. Generally they are quite brittle and, with the exception of one type, must be cast and ground to final size. One of the first alloys of this type to be used commercially contains 71 per cent iron, 12 per cent cobalt, and 17 per cent molybdenum; it is called *Remalloy* or *Comol.* This alloy may be hot-rolled and machined to required size like the carbon steels. To give it permanent-magnet qualities, it is quenched from 1200 deg cent and aged at 700 deg cent for about an hour. Like most dispersion-hardening alloys, the properties of Remalloy do not change appreciably with time. It has found use in meters, receivers, and other devices.

Large quantities of dispersion-hardening alloys are made of the iron-nickel-aluminum type to which have been added cobalt, copper, or titanium. All these alloys, called in this country the *Alnicos*, are brittle and must be cast and ground to size. The addition of titanium up to 8 per cent is sometimes effective in causing high coercive force. Heat treatments and properties are given in Table 2.

Types Heat-treated in a Magnetic Field. The most important of the dispersion-hardening alloys was developed in Holland as *Ticonal* and, with slight variations, is known in the United States as *Alnico 5* and in England as *Alcomax.* The alloy of the proper composition as given in the table is heated to 1300 deg cent and then cooled in air in the presence of a strong magnetic field which must be applied to the magnet in the direction in which the best properties are desired. After aging at about 600 deg cent energy products as high as 5×10^6 are obtained.

Table 2. Typical Properties for Some of the Important Permanent-magnet Materials

Material	Approximate Composition (per cent) Balance Fe	Typical Heat Treatment [1]	Coercivity, H_c oersteds	Retentivity, B_r gausses	Energy Product, (BH)max $\times 10^{-6}$	Form Supplied [2]	Mechanical Nature	Resistivity, ρ micro-ohm-cm	Density, gm/cm^3
C-Mn steel	1Mn 0.9C	Q800	50	10,000	0.2	HR	M	20	7.8
3.5% Cr steel	3.5Cr, 9C 0.3Mn	Q830	65	9,700	0.3	HR	M	38	7.7
6% W steel	6W 0.3Mn 0.7C	Q850	60	10,800	0.3	HR	M	30	8.1
17% Co steel	17Co 0.75C 0.3Mn 2.5Cr 8W	....	150	9,500	0.65	HR	M	....	7.9
36% Co steel	36Co 0.7C 0.3Mn 4Cr 5W	Q950	240	9,500	0.97	HR	M	27	8.3
Remalloy * or Comol *	17Mo 12Co	Q1,200 B700	250	10,500	1.1	HR	M	....	8.4
Alnico * 1	12Al 21Ni 5Co	A1,200 B700	440	7,200	1.4	Cast	G	75	6.9
Alnico * 2	10Al 17Ni 12.5Co 6Cu	A1,200 B600	550	7,200	1.6	Cast	G	65	7.1
Sintered Alnico * 2	10Al 17Ni 12.5Co 6Cu	A1,300	520	6,900	1.4	Sintered	G	....	6.9
Alnico * 4	12Al 28Ni 5Co	Q1,200 B650	700	5,500	1.3	Cast	G	75	7.0
Alnico * 5	8Al 14Ni 24Co 3Cu	AF1,300 B600	550	12,500	4.5	Cast	G	47	7.4
Alnico * 12	6Al 18Ni 35Co 8Ti	....	900	5,700	1.8	Cast	G	....	7.2
Cunife * I	60Co 20Ni	CW B600	550	5,400	1.5	Drawn wire	D	45	8.6
Cunife * II	50Cu 20Ni 2.5Co	CW	260	7,300	0.8	CR	D	....	8.6
Cunico *	50Cu 21Ni 29Co	....	710	3,400	0.85	CR	D	....	8.3
Vicalloy *	52Co 14V	CW 600	400	9,600	2.8	CR	D	....	8.1
"	52Co 9.5V	CW 600	130	14,200	1.35	CR	D	....	8.2
"	53.5Co 8V	600	200	10,000	1.0		D	....	8.2
Powder magnet	Fe(+ Co)	....	580	7,200	1.7	Pressed	M	....	4
Sintered oxide (vectalite) *	$30Fe_2O_3$ $40Fe_3O_4$ $26Co_2O_3$	....	900	1,600	0.5	Sintered	W	10^{12}	2.8
Silmanal *	86.8Ag 8.8Mn 4.4Al	....	595 [3]	575	0.085	CR	M	26	9.0
Pt-Co	77Pt 23Co	Q1,200 B650	2,600	4,500	3.8	Cast	M	50	....

[1] Q = quenched in oil (or water) from temperature (°C) indicated.
B = baked (aged) at temperature indicated.
A = air quenched from temperature indicated.
F = magnetic field applied during cooling.
CW = cold worked to produce properties.

[2] HR = hot rolled.
CR = cold rolled.
M = machinable.
G = must be ground.
D = ductile.
W = weak.

[3] Value given is for $B = 0$; for $B - H = 0$, $H_c = 6{,}000$.

* Trade names.

The interesting oxide magnets made by the Japanese also require a strong magnetic field for developing their best properties. The proper amounts of iron and cobalt oxides (Fe_2O_3, Fe_3O_4, and Co_2O_3) are pressed and heated to 1000 deg cent. They are then cooled and crushed and pressed to final form, then reheated to 1000 deg cent, and finally cooled in a strong magnetic field. Large variations are found in coercivity and retentivity, depending on how the oxides are mixed and treated; H_c ranges from 600 to 1000 and B_r from 4000 to 1500. This material is unusually light in weight and has exceptionally high resistivity. One product, marketed under the trade name *Vectolite*, has a coercivity of 900 and a retentivity of 1600.

Ductile Alloys. Within the last few years a number of ductile permanent-magnet alloys have appeared on the market. The largest application has been in wire and tape form for the magnetic recording of speech. The first alloys of this type were based on the German material containing 20 per cent iron, 20 per cent nickel, and 60 per cent copper, known in this country as *Cunife*. Another variation, called *Cunico*, contains cobalt. Most of these alloys can be cold-drawn to fine wires; in fact, such cold reduction is often necessary to develop their best properties. *Vicalloy*, made of iron, cobalt, and vanadium, is another alloy of this type. Its properties also depend upon the amount of cold reduction, and an energy product of 1.5×10^6 can be developed after cold rolling and annealing at about 700 deg cent. By drastic cold reductions energy products as high as 4×10^6 can be obtained.

Powdered Materials. Several of the brittle alloys mentioned above can be produced from powders which are pressed into the desired shape, sintered at a high temperature, and heat-treated to give the best magnetic properties. This is often advantageous in producing small magnets, when the methods of powder metallurgy can be used. Alnico 2 is important in this class, and Alnico 5 has been used on an experimental scale.

A recent development is material obtained by pressing powder of very fine size, such as that produced by the reduction of iron compounds at low temperature. One product, manufactured in France, is composed essentially of fine particles of iron, sometimes with admixture of cobalt, and has properties much like those of Alnico 2.

Special Alloys. A few other alloys are worthy of brief mention. These are the cobalt-platinum alloys, the manganese-silver-aluminum alloy (*Silmanal*), and certain alloys formed by electrodeposition on a mercury surface, e.g., iron with a small admixture of zinc or iron-cobalt-nickel-aluminum alloys. Experimentally this alloy has been made with coercivities of 400 to 500 and retentivities of 9000 to 11,000.

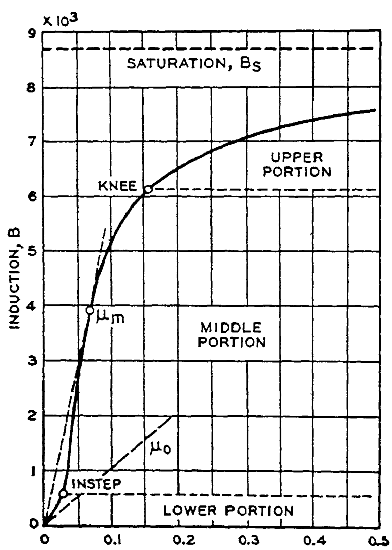


FIG. 12. Magnetization Curve Showing the Points of Special Interest, and Division into Three Main Parts

11. MAGNETIZATION CURVE

In this section will be considered briefly the nature of the changes in magnetization that correspond to different parts of the magnetization curve, and some relations valid in each part.

The magnetization curve may be divided into three parts separated by the "instep" and "knee" (see Fig. 12). At the instep the curve takes a sudden upward turn and the permeability increases rapidly. At the knee this trend is reversed and the curve becomes more and more horizontal and approaches asymptotically to saturation. In each of the three parts of the curve, magnetization proceeds by a different mechanism, as described below.

A magnetic material is composed of many small magnetized regions or *domains*, each of which is always magnetized to saturation in some one direction. When the material as a whole is unmagnetized, these domains are arranged in various directions so that the net magnetization of the material is zero. The effect of the field is to change either the direction in which the domains are magnetized or to change the volume of some of the domains at the expense of their neighbors. This may be made clear by reference to Fig. 13, where the domains are represented by arrows indicating the directions in which they are magnetized. The directions of stable magnetization are determined in an annealed material by the magnetic properties of the crystals of which it is composed, and in severely cold-worked material by the internal strains present.

In the first part of the magnetization curve, below the instep, magnetization proceeds by small displacements of the boundaries between domains, a process illustrated in part (*b*) of Fig. 13. In this portion the permeability usually increases linearly with magnetizing force:

$$\mu = a + bH$$

In the second part of the curve, between the instep and the knee, the domains change direction suddenly and the magnetization changes from one direction of stable magnetization to another. In this section Steinmetz' law of hysteresis is applicable:

$$W_h = \eta B^{1.6}$$

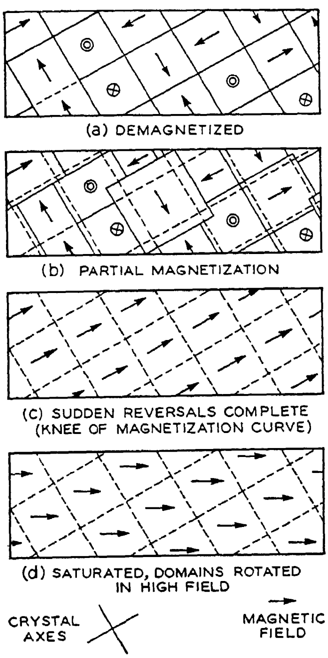


FIG. 13. Diagram Illustrating Changes in Domain Structure with Magnetization in a Single Crystal of Iron

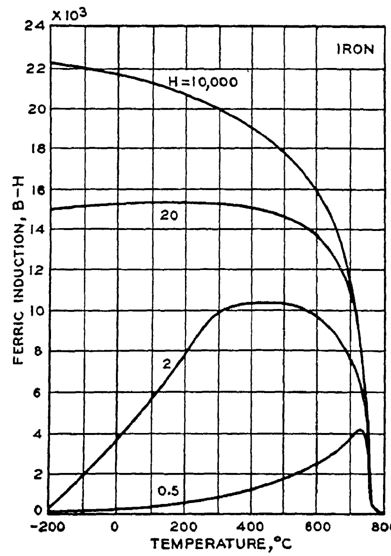


FIG. 14. The Magnetic Induction of Iron Measured at Various Temperatures, with Various Impressed Fields

In the third section the domains rotate smoothly from the stable directions indicated in (*c*) into parallelism with the magnetic field as indicated in (*d*); here the Fröhlich-Kennelly relation

$$\frac{1}{\mu} = c + dH$$

is approximately valid.

Evidence for the domain structure of materials is found in the Barkhausen effect, which proves that sudden changes in magnetization occur in the middle section of the magnetization curve, and in the existence of powder patterns which can be seen under a microscope and which show that the magnetic field at the surface of a demagnetized magnetic material varies from place to place over regions about 0.1 mm apart.

12. EFFECT OF TEMPERATURE

The magnetization of a material can be altered not only by changing the magnetic field but also by varying mechanical stress or temperature. Temperature affects the magnetic properties of all materials, in a way that depends on the induction and the character of the material. Figure 14 shows this effect for iron. Eventually, as the temperature is raised, the material becomes non-magnetic, the temperature at which this occurs being called the *Curie point* of the material. When a high constant magnetic field is present, the magnetization decreases continually as the temperature increases, and at a faster and faster rate as the Curie point is approached. When a low field is present, the permeability first increases with temperature and then decreases again and approaches 1 at the Curie point.

For some applications it is desirable to have a material with a permeability that decreases rapidly as the temperature increases. These materials are used in compensating permanent magnets for changing temperature and for stabilizing pressed powdered cores to make their inductance independent of temperature. The alloys used for this purpose have a Curie point near room temperature so that the material loses its magnetism rapidly as the temperature increases in this region. One such alloy contains about 30 per cent nickel and the rest iron; it has been used for compensating permanent magnets. Another type, containing 80 per cent nickel, 12.5 per cent molybdenum, and the rest iron, is commonly used for stabilizing pressed powdered Permalloy cores. Still another type contains about 35 per cent nickel, 5 per cent chromium, 0.3 per cent silicon, and the rest iron.

13. STRESS AND MAGNETOSTRICTION

Figure 15 shows the way in which the magnetization may be affected by the application of tensile stress. A tension well within the elastic limit of the material will *increase* the magnetization of *45 Permalloy and decrease* that of pure nickel. Under certain conditions the permeability of some materials is increased by a factor of 50 by a stress within the elastic limit. Materials whose permeabilities are increased by tension are said to have positive magnetostriction, because they expand a few parts per million when they are

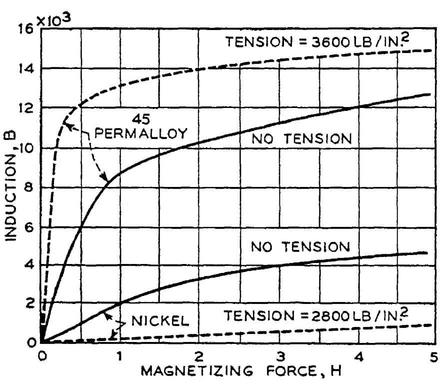


FIG. 15. Magnetization Curves for 45 Permalloy and Nickel as Affected by Tension

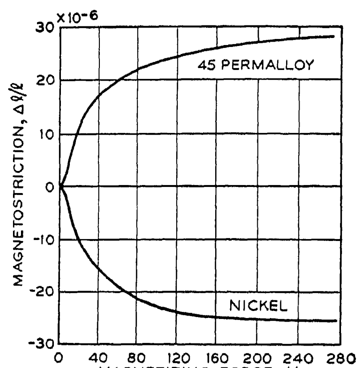


FIG. 16. Fractional Change in Length with Magnetizing Force for 45 Permalloy (Positive Magnetostriction) and for Nickel (Negative Magnetostriction)

magnetized; conversely, the permeability is decreased by tension if magnetization causes contraction of the material (negative magnetostriction, as in nickel). Some materials, like iron, have positive magnetostriction in low fields and negative magnetostriction in high fields.

Figure 16 shows how *45 Permalloy* and nickel change in length as the field strength increases. Such *magnetostriction* is capable of converting magnetic energy into mechanical energy, and nickel is often used in magnetostriction oscillators to produce supersonic vibrations in air or under water, where they are effective in sound ranging. The Japanese have used a new alloy, *Alfer*, containing 13 per cent aluminum and the rest iron, for magnetostriction oscillators; its magnetostriction is about the same in magnitude as that of nickel, but is opposite in sign.

14. EFFECT OF FREQUENCY

When the field acting on a magnetic material is alternated rapidly, eddy currents are induced in the material. They act so as to keep the a-c field from penetrating effectively more than a certain distance below the surface; this distance is measured in a rough way by the expression:

$$s = \frac{\sqrt{\rho/\mu f}}{2\pi}$$

wherein μ = true permeability (d-c).
ρ = 10^3 times resistivity in microhm-centimeters.
f = frequency in cycles per second.

The properties of the material are affected in three ways: (1) the effective permeability is reduced, (2) the energy loss in the material is increased, and (3) there is a time lag between the magnetizing force and the corresponding induction.

Only under certain restricted conditions can the effects of alternating current be predicted with any assurance of correctness. When the permeability is constant, as it usually is only when B *is very small*, the effective permeability, $\bar{\mu}$ (as determined with an inductance bridge) is given by the relation:

$$\frac{\bar{\mu}}{\mu} = \frac{1}{\theta} \cdot \frac{\sinh\theta + \sin\theta}{\cosh\theta + \cos\theta}$$

in which $\theta = 2\pi t\sqrt{\mu f/\rho} = t/s$.
t = thickness of sheet in centimeters.

Effective permeabilities, calculated by means of this equation, are compared with actual measurements in Fig. 17.

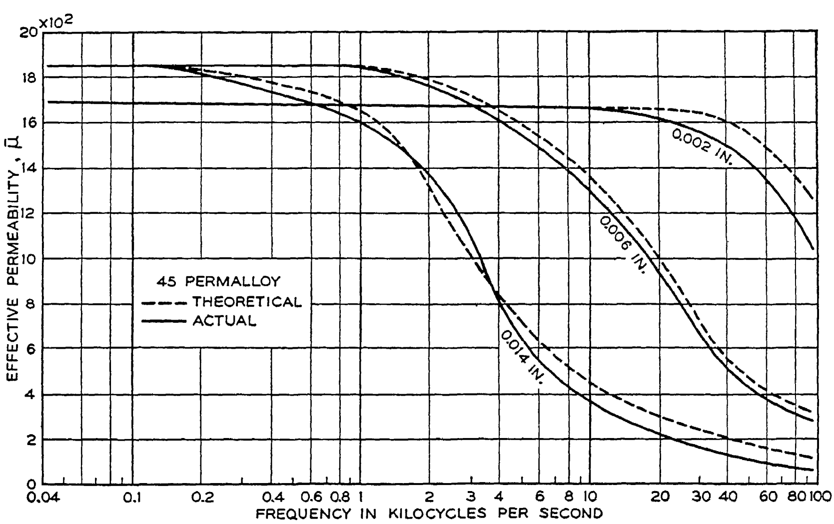


Fig. 17. Change in Effective Permeability with Frequency for 45 Permalloy of Different Thicknesses

When $\theta \gg 1$ (at high frequencies) this expression reduces to:

$$\bar{\mu} = \frac{\sqrt{\mu p}}{2\pi t} \cdot \frac{1}{\sqrt{f}} = \frac{\mu s}{t}.$$

Similar expressions are applicable to specimens in the form of wire or cylinders.

When μ is constant and $\theta < 1$ (frequency and induction low) the power loss may be expressed by the equation:

$$\frac{R}{L} = 2\pi f(h_0 + hB + ef)$$

in which R = excess of a-c resistance (by a-c bridge) over d-c resistance, in ohms.
L = inductance of coil in henrys.
B = maximum induction in gausses.

This relation is especially adapted to materials used in communication circuits in which there are feeble alternating currents. The constants h and e measure the hysteresis and eddy-current losses, respectively, and h_0 is of unknown origin, important only at the lowest B's.

At low frequencies ($\theta < 1$) and *high inductions* (B = 1000 to 14,000 in silicon-iron), the power loss in ergs per centimeter^3 per second is:

$$W = \eta B^{1.6} f + e B^2 f^2$$

The hysteresis constant, η, and the eddy-current constant, e, can be determined in an approximate way by plotting W/f vs. f for given values of B. At high frequencies eddy-current loss is usually more important than hysteresis loss, and is given in ergs per centimeter^3 per second by the relation:

$$W_e = \frac{\pi t \bar{B}^2 f^{3/2}}{\sqrt{\mu\rho}}$$

in which $\bar{B}$ is the value of B averaged over the cross-section of the sheet.

At frequencies higher than 10^8 cps the true permeability, μ, of magnetic materials begins to decrease substantially, approaching a value of the order of 1 at frequencies around 10^9 to 10^{11} cps.

Magnetization also affects the resistivity of magnetic materials. The change is almost invariably an increase in resistivity with magnetization, the amount of the increase varying from less than 1 per cent to about 5 or 10 per cent at room temperature, and even more at low temperatures. Similarly Young's modulus may be changed by about 10 per cent by magnetizing to saturation. Unusual variations are also observed in specific heat, thermal expansion, and other physical properties of magnetic materials.

15. MEASUREMENT OF MAGNETIC CHARACTERISTICS

Although it requires many different measurements to determine all the magnetic characteristics of a material, the most important properties can be obtained from a magnetization curve and hysteresis loop, an a-c measurement of the permeability and losses, and a measurement of the incremental permeability at various polarizing inductions.

Of the several methods that can be used for measuring magnetic properties, the *ballistic ring test*, due to Rowland, is perhaps the most reliable. In this test, a ring sample is wound with two uniformly distributed windings, consisting of a primary connected to a source of current, and a secondary connected to a *ballistic galvanometer* or *fluxmeter*. The induction produced by current in the primary winding is observed in terms of the fluxmeter deflection as the primary current is changed suddenly or reversed. The use of a ring sample eliminates the possibility of errors due to air gaps. To obtain uniformity of magnetizing force throughout the sample, the ratio of the outside diameter to the inside diameter of the ring should be not greater than 1.2. Figure 18 shows a typical electrical circuit for

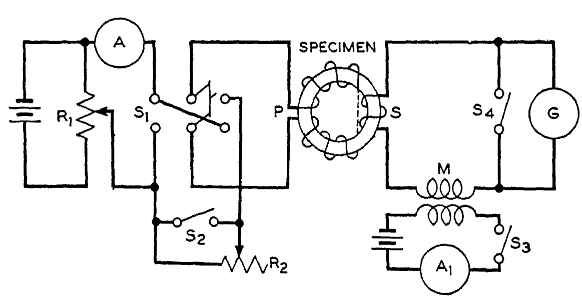


FIG. 18. Simplified Diagram of a "Ballistic Test" Circuit

this type of test, including a mutual inductance for calibrating the galvanometer or fluxmeter, G, and resistances R and switches S for regulating and changing the current in the primary winding, P. The field is calculated using the relation:

$$H = \frac{4N_pI}{10d}$$

and the induction using the expression:

$$B = K\delta = \frac{I_c M \times 10^8 \delta}{\delta_c N_s A_s}$$

in these equations N_p = number of turns in primary winding.
I = primary current in amperes.
d = mean diameter of ring in centimeters.
I_c = calibrating current in amperes.
M = calibrating mutual inductance in henrys.
δ_c = calibrating deflection resulting from a reversal of I_c.
N_s = number of turns in secondary winding.
A_s = sectional area of sample in square centimeters.
δ = fluxmeter deflection resulting from reversal of current in primary winding.

Straight bar or rod samples are sometimes tested with this circuit. A long solenoid is used for producing the magnetizing force, and the secondary winding or search coil is placed around a short central portion of the sample. However, under these circumstances the true magnetizing force is difficult to determine because the field from the magnetic poles produced at the ends of the sample reacts with the field of the solenoid. The field created by the sample itself is sometimes called the end effect or demagnetizing field. Its value is usually specified by the *demagnetizing factor*, N, which depends on the ratio length/diameter of the rod. The field, H, acting at the center of the rod is the resultant of the field in the solenoid, H_0, and the demagnetizing field:

$$H = H_0 - \frac{N}{4\pi}(B - H)$$

The apparent permeability, μ', is given by B/H_0, and its relation to the true permeability, μ is given by:

$$\frac{1}{\mu} = \frac{1}{\mu'} - \frac{N}{4\pi}$$

The relation between μ and μ' for cylindrical rods is shown graphically in Fig. 19.

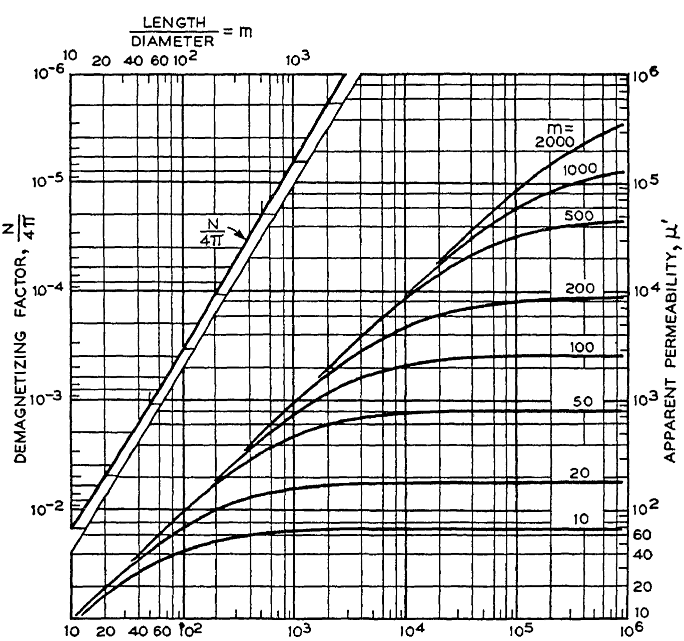


FIG. 19. Relation of the Apparent Permeability to the True Permeability for Cylindrical Rods of Various Ratios, m, of Length to Diameter. Also, Demagnetizing Factors, $N/4\pi$, as Dependent on m.

Various types of permeameters are also used with the circuit shown in Fig. 18. They are especially useful for measuring permanent-magnet materials. Permeameters usually are designed to test straight bar samples clamped against a yoke of very low reluctance. Measurement of samples of high permeability, so tested, are subject to error due principally to the effects of the air gaps in the circuit. Only a few of the many types will be described.

The *Fahy permeameter* is commonly used for testing materials like iron and silicon-iron as well as some of the magnet steels of relatively low coercive force. It is suitable for tests at magnetizing forces up to 300 oersteds. This instrument, shown in Fig. 20, has one large magnetizing winding on a yoke of silicon-iron. Pole pieces extending from either end of the yoke are arranged so that bar samples can be clamped to them. The magnetizing force is measured by an air-core solenoid (*H* coil) mounted across the ends of the pole pieces and above the sample. A winding enclosing the sample acts as a secondary (*B* coil) and measures the induction with the aid of a galvanometer as in Fig. 18.

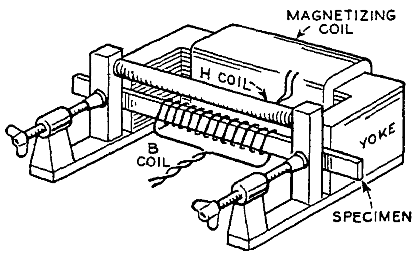


FIG. 20. Descriptive Drawing of the Fahy Permeameter

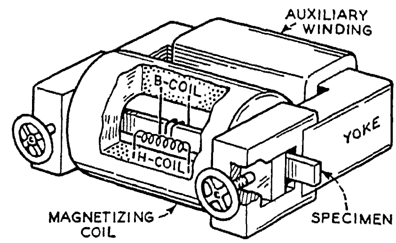


FIG. 21. Descriptive Drawing of the Babbitt Permeameter

The *Babbitt permeameter* (see Fig. 21) can be used for testing at magnetizing forces as high as 1000 oersteds. Several different high-permeability materials are used in the yoke to give low reluctance over a wide range of magnetizing force, and the magnetizing coil encloses the sample instead of being wound on the yoke alone as in the Fahy. Smaller windings are placed on the yoke to compensate for the reluctance of the air gaps. Magnetizing force and induction are measured with *H* and *B* coils placed near and around the middle of the sample.

More accurate tests of high-permeability materials can be made by means of the *Burrows permeameter*. This type requires two samples clamped between two connecting yokes of high-permeability material completing the magnetic circuit. In addition to the magnetizing coils around the samples, there are compensating coils around each end of the samples to give more adequate corrections for the effect of the air gaps at the joints. *B* coils are wound around the middle of each sample, and two search coils are placed on either side of each *B* coil. With proper adjustment the conditions of test more nearly approach those in the ring test, but the test is time-consuming because of the number of adjustments required.

Small air gaps are not very important when testing permanent-magnet alloys, but modern magnet materials require permeameters that can produce very high magnetizing fields. The saturation permeameter and the high-*H* permeameter are frequently used for this purpose.

The *saturation permeameter* is very similar to the Babbitt permeameter except that the magnetizing coil is larger and artificially cooled, and no compensating coils are used on the yoke. Magnetizing forces as high as 2500 oersteds are readily obtained with this instrument.

The *high-H-permeameter* developed by the Bureau of Standards can produce even more powerful fields and can be used to test any of the modern permanent-magnet alloys. In this permeameter four large coils are used. Two are wound on the yokes and two on the pole pieces clamped to the specimen (see Fig. 22). The induction is measured in the usual way with a coil wound around the center of the sample. A small rotatable *H* coil is arranged so that readings can be taken at different distances from the surface of the sample; the data so obtained give a curve of the variation in *H* and indicate by extrapolation the true value of field at the center of the sample.

In practice, magnetic materials are often subjected to alternating fields, and it is, therefore, important to measure magnetic permeability and energy losses by a-c methods. These include the use of a-c bridges, wattmeters, cathode-ray oscilloscopes, and various other instruments, a few of which will be described.

One of the common bridge circuits is the *Maxwell bridge.* The simplest form of this bridge consists of a pair of resistances for ratio arms, a variable resistor and variable inductance to balance the impedance of the sample, and a detector which may be a galvanometer, a sensitive voltmeter, or a telephone receiver (see Fig. 23). The bridge is useful for measuring apparent permeabilities and losses for low inductions at frequencies in the audio range. It is not suitable for testing at high inductions because of errors introduced by wave-form distortion.

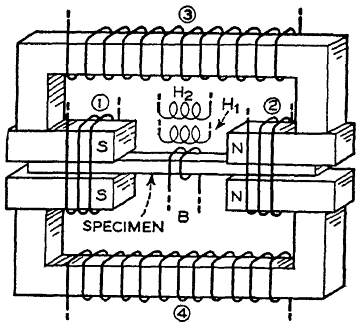


FIG. 22. Schematic Diagram of the High-H Permeameter

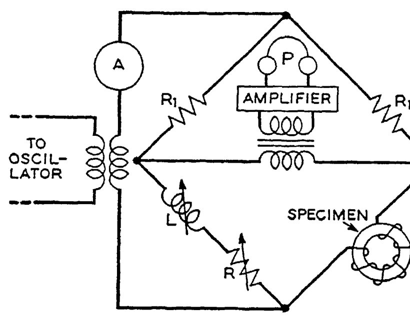


FIG. 23. Simplified Maxwell Bridge Circuit for Determining Equivalent Series Resistance, R, and Inductance, L, of a Coil Containing a Specimen of Magnetic Material

Incremental or superposed permeability measurements can also be made with the Maxwell bridge, by using an additional winding on the sample connected in series with a large inductance and a source of direct current. The inductance in the d-c circuit must be large enough to keep the alternating current in this circuit at a negligible value. Incremental permeability is important in the transformers of vacuum-tube amplifiers and in polarized apparatus such as telephone receivers where the material is subjected to both d-c and a-c fields (see Fig. 24).

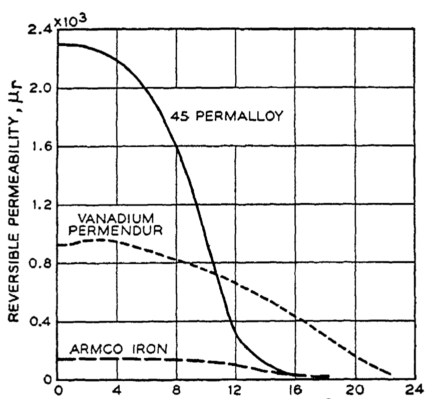


FIG. 24. Change in Reversible Permeability with Polarizing Induction for Several Materials

Properties are often determined at frequencies above the audio range with a-c bridges of the resistor-capacitor type, and Q-meters. For further information on a-c bridge methods, see Measurement of Inductance and Effective Resistance, Section 11.

For magnetic materials in sheet form, it is convenient to test samples made of sheared strips. The *Epstein test,* in common use for testing the core loss of materials such as silicon-iron sheet, uses samples of this form. Primary and secondary windings are placed on four hollow square forms mounted in the form of a square. The primary exciting current is measured with an a-c ammeter, A, and the induction is indicated by an "average" voltmeter, sometimes called a *flux* voltmeter, connected across the secondary as indicated in Fig. 25. The core loss is determined from the reading of a wattmeter, W. The test strips used in the Epstein are usually stacked with overlapping joints for permeability tests but may be stacked with butt joints for core-loss tests. This method gives reliable data up to

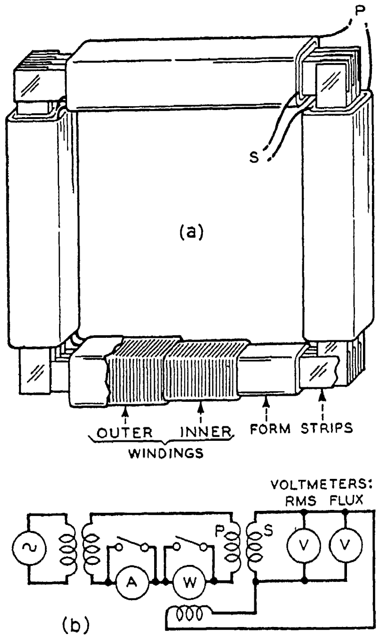


FIG. 25. Descriptive Drawing and Simplified Circuit Diagram for the Epstein Test for Determining Watt Loss

high inductions, and it is particularly useful for studying grain-oriented materials. For details of this test and other test methods, the latest issue of ASTM specifications should be consulted.

Cathode-ray oscilloscopes are sometimes used to give rapid indications of the a-c properties of materials. By means of a simple integrating circuit, hysteresis loops can be produced on the screen. A simplified circuit of this type is shown in Fig. 26. This test is

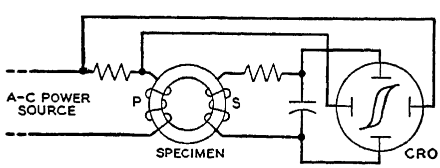


FIG. 26. Simplified Circuit for Producing Hysteresis Loops on Cathode-ray Screen

not as precise as those described above, but because of its rapidity it finds frequent application in certain types of production testing.

BIBLIOGRAPHY

ASTM Spec. A-34.
Altman, G. O., and H. Beller, Radio Frequency Cores of High Permeability, *Electronic Industries*, Vol. 4, 86 (1945).
Bozorth, R. M., Present Status of Ferromagnetic Theory, *Elec. Eng.*, Vol. 54, 1251 (1935).
The Physical Basis of Ferromagnetism, *Bell Sys. Tech. J.*, Vol. 19, 1 (1940).
Becker, R., and W. Döring, *Ferromagnetism* (in German). Springer, Berlin (1939).
Cioffi, P. P., Hydrogenized Iron, *Phys. Rev.*, Vol. 39, 363 (1932).
Dillinger, J. F., and R. M. Bozorth, Heat Treatment of Magnetic Materials in a Magnetic Field, *Physics*, Vol. 6, 279 (1935).

Electrical Engineering Staff of M. I. T., *Magnetic Circuits and Transformers.* John Wiley, New York (1944).
Ellis, W. C., and E. E. Schumacher, A Survey of Magnetic Materials in Relation to Structure, *Bell Sys. Tech. J.*, Vol. 14, 8 (1935).
Elmen, G. W., Magnetic Alloys of Iron, Nickel and Cobalt, *Elec. Eng.*, Vol. 24, 1292 (1935).
Hornfeck, A. J., and R. F. Edgar, The Output and Optimum Design of Permanent Magnets Subjected to Demagnetizing Forces, *A.I.E.E. Trans.*, Vol. 50, 1017 (1940).
Legg, V. E., Survey of Magnetic Materials and Applications in Telephone System, *Bell Sys. Tech. J.*, Vol. 18, 438 (1939).
Legg, V. E., and F. J. Given, Compressed Powder Molybdenum Permalloy for High Quality Inductance Coils, *Bell Sys. Tech. J.*, Vol. 19, 385 (1940).
Messkin, W. S., and A. Kussmann, *Ferromagnetische Legierungen* (in German). Springer, Berlin (1932).
Rotors, H. C., *Electromagnetic Devices.* John Wiley, New York (1941).
Ruder, W. E., Permanent Magnet Steels, *Iron Age*, Vol. 157, 65 (1946).
Sanford, R. L., An Apparatus for Magnetic Testing at Magnetizing Forces up to 5000 Oersteds, *Bur. Standards J. Research*, Vol. 23 (1939).
Permanent Magnets, *Bur. Standards Circ.* C448 (1944).
Spooner, T., *Properties and Testing of Magnetic Materials.* McGraw-Hill, New York (1927).
Woldman, N. E., and R. J. Metzler, *Engineering Alloys.* Am. Soc. Metals (1945).
Yensen, T. D., Magnetically Soft Materials, *Trans. Am. Soc. Metals*, Vol. 27, 797 (1937).

SECTION 3

RESISTORS, INDUCTORS, CAPACITORS

RESISTORS, INDUCTORS, CAPACITORS

RESISTORS AND RHEOSTATS

By Paul S. Darnell and Arthur H. Schafer

1. GENERAL

DEFINITIONS. A *resistor* is a circuit element whose primary function is to introduce electric resistance into an electric circuit.

A *rheostat* is an adjustable resistor which is provided with mechanical means for changing its resistance value without opening the circuit in which it may be connected. Its primary function is to adjust the current in a circuit or portion of a circuit in which it is connected. It may be in the form of a three-terminal potentiometer as defined below, or it may have only two terminals.

A potentiometer is defined by American Standards Association as a measuring instrument by means of which an electromotive force in one of the arms of the circuit may be measured in terms of one or more other electromotive forces and the constants of the potentiometer circuit. However, the term potentiometer is most commonly used to refer to any adjustable resistor having three terminals, two of which are connected to the ends of the resistance element and the third to a contact which traverses the resistance element without discontinuity. Its primary function is to convert an impressed voltage into a source of voltage which can be adjusted from a small percentage of the impressed voltage to approximately the magnitude of the impressed voltage.

PHYSICAL AND ELECTRICAL CONSIDERATIONS. Numerous factors, which will determine the type as well as the physical size and shape, must be considered in the design or selection of the proper resistor or rheostat for a specific application. The most important of these are: (1) resistance value and tolerance; (2) power dissipation under normal and trouble operating conditions; (3) frequency characteristic, phase angle, and change in resistance over a frequency range; (4) stability with age and changing conditions of temperature, humidity, and voltage; (5) mounting arrangements—space and shape requirements; and (6) relative cost of available resistor structures which will fulfill the circuit requirements in whole or in part.

Resistance Value. In general two classes of material are used for resistors: (1) pure metals and metal alloys in which resistance value for any given metal is determined largely by its physical dimensions, and (2) a composition or mixture of a carbon or metallic conductor with an insulating material in which the resistance value is determined by the relative proportion of conductor and insulator.

In the first group the element is usually a wire or strip having relatively low resistivity, which places a definite limitation on the resistance value that can be achieved for a given volume. The advantages are high stability with age, low temperature coefficient, low voltage coefficient, and low microphonic noise level. Its disadvantages are relatively high cost for higher resistance values, corrosion hazard under adverse conditions of voltage and humidity, limitation on high resistance values, and usually poor frequency characteristics at higher frequencies. Resistance range is normally from about 0.1 ohm to 1 megohm. In addition to the wire-wound resistors in this class are the metallic or carbon precision film-type resistors (to be discussed later) which circumvent some of the disadvantages enumerated here, since they permit attainment of high resistance in small space and have greatly improved frequency characteristics.

In the second group of resistors, since resistance value becomes a matter of composition of a mixture, the entire range of resistors in general use (10 ohms to 22 megohms) can be made in the same physical form and volume, so that size is determined largely by power dissipation desired. Other advantages are low cost, improved performance at high frequency, and, when wattage is not a factor, small size, light weight, and ease of mounting. Disadvantages are lesser stability with time, temperature, and humidity compared with wire-wound resistors, broader manufacturing tolerances, and higher noise level.

A third class of resistors, namely varistors and thermistors (see articles 8 to 10), is currently finding wide application as special-purpose circuit elements that are characterized chiefly by their high sensitivity to temperature and voltage and their relatively high resistivity. In this group the elements consist of semiconductor materials in the range between metallic conductors and insulating materials.

Resistance Tolerance. The matter of manufacturing tolerance with respect to the limits of resistance value within which a resistor shall initially be held is important from the standpoint both of the proper functioning of the circuit involved and of the design of the resistor itself. Practical tolerances of low-cost resistors commercially available are of the order of ±5 per cent to ±20 per cent for composition type, and ±1 per cent to ±10 per cent for wire-wound and precision film types.

Of equal importance to the initial manufacturing tolerances is the desired stability of the resistor during its life. To obtain stability of a lesser order of magnitude than the original manufacturing limits can be fairly simple at even the closest limits of the ranges indicated above but will probably become the major problem or even an impossibility for units adjusted to initial tolerances of one-quarter to one-tenth of those figures, for example.

ENERGY DISSIPATION—TEMPERATURE RISE—POWER RATINGS. Since normally the total electrical energy supplied to a resistor is dissipated in the form of heat, the resultant temperature rise will, under adverse conditions, constitute a potential hazard both to the resistor itself and to the materials of surrounding objects. Industrial Control Standards (NEMA, July 1, 1946, IC4-22) state that the temperature rise for bare resistive conductors shall not exceed 375 deg cent as measured by a thermocouple in contact with the hottest spot on the bare conductor, and for resistive conductors imbedded shall not exceed 300 deg cent as measured by a thermocouple in contact with the hottest spot on the surface of the imbedding material. The establishment of such maximum power ratings is possible only for power-type resistors constructed entirely of inorganic materials not adversely affected either physically or chemically by the heat generated. Ratings for other resistors are normally established on the basis of the maximum temperature at which they can operate continuously without deterioration of their performance or their component parts. Since power ratings under the various standards are predicated on operation in still air and free space at an ambient temperature of the order of 25 deg cent, a status seldom realized under actual operating conditions, good engineering practice dictates derating the resistor to reflect the specific conditions of use. In the absence of exact data to indicate the amount of derating necessary, a figure of 50 per cent is commonly applied. In addition to consideration of the normal wattage at which the resistor will be required to operate, it is frequently advisable to determine the power the resistor must dissipate under predictable abnormal or trouble conditions, and when possible, to select a resistor that will operate safely under such conditions as well.

FREQUENCY CHARACTERISTIC. In considering the behavior of a resistor over a frequency range, it must be remembered that in any practical design in which the element has finite dimensions it will necessarily contain components of all three parameters (Fig. 1—R, L, and C) and will approach the characteristics of a pure resistor over a limited frequency range. The problem of design is to approach a pure resistance with the inductance and capacitance regarded as parasitic values, useful only to the extent to which they can be made to nullify each other over the frequency range in which the resistor is to be used. A typical resistor may be represented as a two-terminal network comprising three elements as shown in Fig. 1.

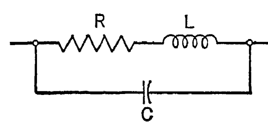


FIG. 1. Equivalent Network of a Resistor Having Inductance and Distributed Capacitance

The impedance Z of this circuit may be expressed in terms of effective inductance L' and effective resistance R' of the system as follows: $Z = R' + j\omega L'$. It can readily be shown that in terms R, L, and C the following relations hold:

$$R' = \frac{R}{1 - \omega^2 C(2L - CR^2) + \omega^4 C^2 L^2} \tag{1}$$

$$L' = \frac{L - CR^2 - \omega^2 CL^2}{1 - \omega^2 C(2L - CR^2) + \omega^4 C^2 L^2} \tag{2}$$

Also the phase angle $\Phi = \tan^{-1}(\omega L'/R')$, from which

$$\text{Tan } \Phi = \frac{\omega(L - CR^2 - \omega^2 CL^2)}{R} \tag{3}$$

Since in a resistor both L and C are small, eqs. (1), (2), and (3) may be rewritten:

$$R' = R[1 + \omega^2 C(2L - CR^2)] \tag{4}$$

$$L' = L - CR^2 \tag{5}$$

$$\text{Tan } \Phi = \frac{\omega(L - CR^2)}{R} \tag{6}$$

It will be observed that the phase angle will be zero if $L - CR^2 = 0$, but that the change in resistance will be zero only if $C = 0$ or if $2L - CR^2 = 0$. It follows then that both the phase angle and the change in resistance with frequency can be zero simultaneously only for the conditions $C = 0$, $L = 0$. These are important relations, for it means that a resistor may be designed with a small or zero phase angle over a considerable frequency range merely by keeping to a minimum the quantities L/R and CR and making them as nearly equal as is warranted for the particular design under consideration.

A second factor causing change of resistance with frequency is the so-called skin effect, which arises from the fact that elements or filaments of current at different points in the cross-section of the conductor encounter different components of inductance. This is because of the greater mutual inductance between elements at the center of the conductor. These unequal inductances over the cross-section tend to produce unequal current densities, with the minimum at the axis and the maximum at the periphery. Such unequal current densities reduce the effective cross-section of the conductor and tend to increase its effective resistance. Since the determination of skin effect by computation is involved, it is sufficient to observe that this effect may be kept to the minimum by (1) using a conductor of small cross-section, (2) using a conductor of high specific resistance, and (3) for a wire-wound resistor, winding the wire to have the minimum effective inductance. It will be seen that these are also conditions for other good characteristics in an alternating-current resistor. (For a more detailed description of skin effect and methods of computation, see *National Bureau of Standards Circular* 74.)

Small Phase Angle over a Frequency Range. As stated previously, to insure small phase angle over a given frequency range requires that residual inductance and capacitance be held to the minimum, zero phase angle being realized when one is made to compensate the other. Such compensation is achieved by satisfying the equation $L = CR^2$. The major problem for low resistance is the reduction of inductance; for high resistance it is the reduction of capacitance. In actual practice it is found that for values up to about 100 ohms the problem is one of inductance; for values over several thousand ohms the problem is mostly capacitance; for values in between, both capacitance and inductance must be considered. Of course these figures show merely the broad boundaries, and considerable variation will be encountered due to differences in structure, type of winding, and capacitance to ground.

STABILITY WITH AGE, TEMPERATURE, AND HUMIDITY. *Aging* may be defined as any permanent change in resistance value with time, measured under the same conditions of temperature and humidity. It is generally caused by strains set up in the resistor element, either in the original manufacturing process as when wire is wound tightly on a core or when the element is pressed or molded into shape, or during the life of the unit when the supporting structure is distorted as the result of aging of the structure itself. In a composition resistor it is also brought about by chemical and physical changes in the materials of composition, particularly in the deterioration of the insulating materials used in the binder and filler. Strains in the winding may often be relieved by a preaging treatment (generally baking from 1 to 24 hours). To prevent warpage of the structure, and to guard against other effects discussed later, resistors are protected against moisture in various ways. Impregnation with wax, varnish, or asphalt compounds is common practice, and frequently an exterior covering of moisture-resistant material is used. Proper selection of structural materials is probably the most important single factor in aging protection. Metal, which appears to be the ideal material for this purpose, usually cannot be used, partly because of its effect on the residuals, and partly because proximity to the winding would be an added hazard in promoting electrolytic corrosion.

Temperature Variations. Stability with change in temperature is obtained by a proper selection of the resistive material. For composition-type resistors this material is at present limited almost entirely to carbon or graphite for the conducting portion of the mix; a wide variety of insulating materials is used for binder and filler, and frequently the choice is controlled by considerations other than temperature coefficient. For wire-wound resistors the desired temperature characteristics are determined largely by the wire selected.

Humidity Conditions. In addition to warping of structural materials two other humidity effects must be considered in resistance design: the effect on residual inductance and capacitance, and the effect on corrosion. When moisture gets into a resistance structure it causes a change in the dielectric properties of the materials used therein, and a consequent change in the residual capacitance. This will often change the phase angle by an appreciable amount. The danger of corrosion which takes place when impurities and moisture are present is generally the more serious consideration in protection against humidity, especially where fine wires are used.

CLASSIFICATIONS. Resistors have been variously classified in terms of performance, power dissipation, structure, usage, and manufacturing tolerances. Broadly, the following classifications are in general use today and have been adopted, with proper subdivisions, by the various standards agencies: (1) fixed wire-wound, (2) fixed composition type, (3) precision film type, (4) rheostats and potentiometers (wire-wound and composition), (5) special-purpose types (chiefly thermistors and varistors). Each of these categories will be treated separately below.

RESISTOR TESTS AND SPECIFICATIONS. Aside from a marked quality improvement of the resistor product, one of the major benefits derived from the past war was the promulgation and adoption by the armed services, and to a lesser extent by industry, of resistor standards of performance, size, resistance value, and tolerance. These can be found by reference to the following specifications:

JAN-R-11	Resistors, Fixed Composition.
JAN-R-26	Resistors, Fixed Wire-Wound, Power Type.
JAN-R-93	Resistors, Accurate, Fixed, Wire-Wound.
JAN-R-184	Resistors, Fixed, Wire-Wound (Low-power).
JAN-R-22	Rheostats, Wire-Wound, Power Type.
JAN-R-19	Resistors, Variable, Wire-Wound (Low Operating Temperature).

Specifications include tests of both mechanical and electrical qualities, including measurement of strength of leads, strength of resistor, resistance load life, load characteristics, voltage characteristic, temperature coefficient, microphonic noise, and effects of humidity, overload, and aging. In addition the Radio Manufacturers Association is in the process of preparing corresponding specifications for commercial usage, less severe and exacting. (See also Section 11.)

2. WIRE-WOUND RESISTORS

The many varieties of wire-wound resistors in use are classified in numerous ways, some by the manufacturer and some by the application to which the resistor is put. Thus we have power type, high-frequency resistors, precision type, flat type, tubular, plate, ballast, cord type, lead mounting, ferrule type, spool type, sectionalized bobbin type, and many others. As might be expected there is a great deal of overlapping in categories, and the descriptive terms are of course relative and at times misleading. Wire-wound resistors will be arbitrarily subdivided here into the several main classifications currently accepted by most manufacturers and users.

POWER-TYPE RESISTORS. Power resistors consist of that class of resistor whose primary function is to dissipate relatively large amounts of power in a comparatively small space. In this group, materials of construction are selected first because of their ability to withstand heat. Normally temperature rise is limited to 300 deg cent which with 25 deg cent ambient means a hot-spot temperature of 325 deg cent. Where organic materials are used (usually to obtain increased resistance to humidity or to facilitate construction of special features requiring machined or molded parts) temperature rise is limited to 125 deg cent and wattage ratings are reduced to about one-third of maximum values obtainable with complete inorganic construction. Typical construction consists of a cylindrical ceramic-core tube with encircling band terminals at each end, with a single inductively wound layer of wire made of resistance material selected for the qualities desired. Because of corrosion hazards and considerations of mechanical strength, choice of wire size is usually limited to minimum 0.002-in. diameter (0.0025 in. in JAN specifications), although most manufacturers will wind wire down to 0.001 in. or less in diameter if requested to do so. The resistor winding is protected against mechanical injury and the effects of moisture. Electrical insulation is obtained by applying a cover coat of either vitreous enamel or cement. A typical assortment of sizes commercially available in this type is shown in Table 1.

Table 1. Ratings and Dimensions

Nominal Ratings, watts	Range of Resistance Values, ohms	Approximate Core Sizes, in.	
		Diameter	Length
5	0.5 to 1,600	5/16	1
10	0.3 to 5,000	5/16	1 3/4
20	0.3 to 10,000	9/16	2
25	0.3 to 13,000	5/8	2
30	0.5 to 23,000	3/4	3
40	0.6 to 28,000	3/4	3 1/2
50	0.8 to 40,000	3/4	4 1/2
75	1.2 to 50,000	1 1/8	4 1/4
115	1.9 to 90,000	1 1/8	6 1/2
160	2.6 to 120,000	1 1/8	8 1/2
200	3.6 to 150,000	1 1/8	10 1/2

Maximum resistance value is based on use of 0.002-in.-diameter wire. With 0.001-in. wire, values up to eight times these maximum figures may be obtained. Wattage ratings are based on a permissible 300 deg cent temperature rise. Maximum voltage which may be applied is properly held within about 500 volts per inch of length, subject, of course, to permissible power dissipation.

Other varieties of power-type resistors include the following: strip type, plaque type, disk type, plate type, supported ribbon, ferrule or axial terminal, adjustable type,

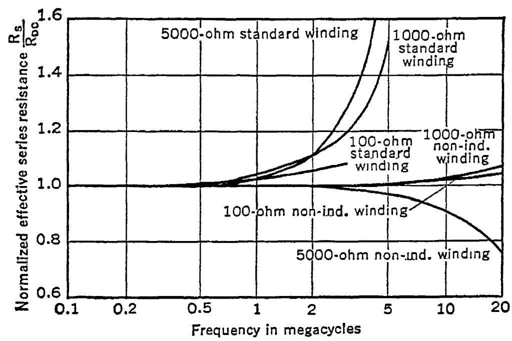


FIG. 2. Resistance Frequency Characteristics of Several Wire-wound Vitreous Enamel Resistors—50 Watts, 9/16 in. Diameter x 4 in. Core (Courtesy Ohmite Mfg. Co.)

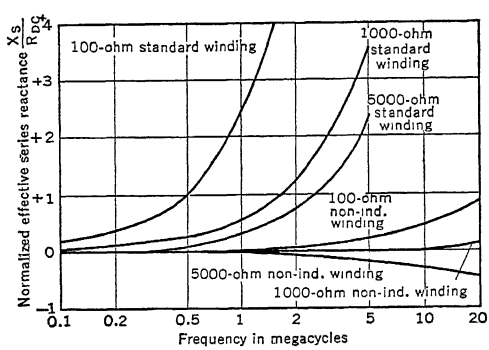


FIG. 3. Reactance Frequency Characteristics of Several Wire-wound Vitreous Enamel Resistors—50 Watts, 9/16 in. Diameter x 4 in. Core (Courtesy Ohmite Mfg. Co.)

molded type, and those with multiple or tapped windings. The adjustable type has a lengthwise strip of winding bared down one side of the core tube, and the resistor is provided with an extra encircling band terminal for clamping down on and tapping the winding at any point along its length.

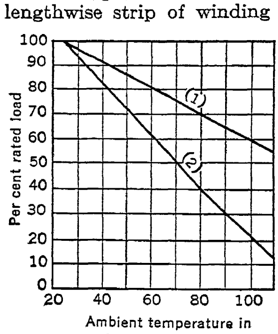


FIG. 4. Resistor Derating for High Ambient Temperatures as Specified in JAN-R-26. Curve 1—Nominal rating based on 250 deg cent temperature rise. Curve 2—Nominal rating based on 125 deg cent temperature rise.

Close-tolerance Resistors. Although usual manufacturing tolerances available are either ±5 or ±10 per cent, closer tolerances down to about ±1 per cent can usually be obtained at increased cost. Tolerances closer than 1 per cent are not generally furnished in power-type resistors since self-heating frequently changes the resistance value by at least 1 per cent. In the case of vitreous enamel resistors the high firing temperatures involved in applying the protective coating and the consequent resistance changes during manufacture make closer adjustment impracticable. Sometimes varnishes, lacquers, or asphalt coatings are substituted for the high-temperature coverings to facilitate initial adjustment, to permit use of low-temperature-coefficient wires without loss of desirable characteristics brought about through additional heat treatment, and to provide further protection against weathering in service. Such special coatings may necessitate derating as much as 80 or 90 per cent.

Layer Windings. Power-type resistors are occasionally wound in layers to achieve a greater range of resistance value. Resistance wire having insulation capable of withstanding high temperatures as well as having a high degree of mechanical strength must be chosen for this type of resistance winding.

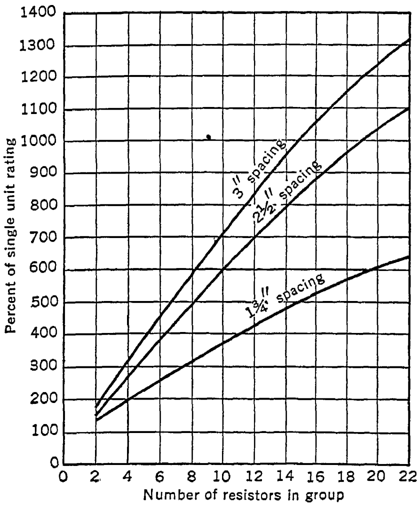


FIG. 5. Resistor Derating When Mounted in Groups. Spacings are centerline to centerline of resistors mounted horizontally—1 1/8 in. diameter core tubes. All resistors of equal length and power rating. (Courtesy Ward Leonard Electric Co.)

Non-inductive Windings. Many of the power-type resistors may be obtained with windings of low inductance. Generally the Ayrton-Perry winding is used (see description under "Precision-type Resistors"), which has the added advantage of low distributed capacitance and greatly increases the frequency range over which the resistor can be used effectively, particularly for low resistance values (up to about 10,000 ohms). Figures 2 and 3 show plots of the normalized effective series resistance and normalized effective series reactance against frequency for both standard and non-inductive (Ayrton-Perry) windings on 50-watt core sizes for resistance values of 100, 1000, 5000 ohms, and are representative of the improved performance that may be expected from this type of winding. Normalized resistance or reactance is to be considered as the ratio of resistance or reactance to the d-c resistance of the resistor.

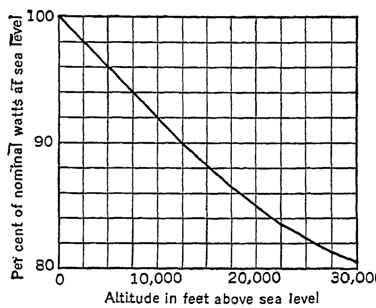


FIG. 6. Resistor Derating When Operated Above Sea Level (Courtesy Ward Leonard Electric Co.)

Typical behavior and characteristics of some power-type resistors are shown in Figs. 4 to 10. Figures 4 to 7 show the required derating of resistors operated under unfavorable conditions; Figs. 8, 9, and 10 show behavior under overload and intermittent duty.

LOW-POWER AND PRECISION-TYPE RESISTORS. This category is made up of those resistors in which power dissipation is a minor consideration; it comprises a wide variety of shapes and sizes. Since power dissipation is not an important factor, many more materials are available for body structures and protective coatings, including plastics and molding compounds, which are more readily adaptable to the form of unit required.

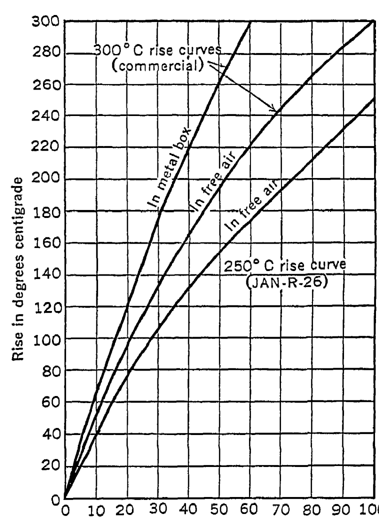


FIG. 7. Temperature Rise of Resistors in Free Air, and Enclosed in a Metal Box. Unit mounted in unventilated metal box reaches maximum permissible temperature when dissipating only about 60 per cent of normal rating. (Courtesy Ward Leonard Electric Co.)

Precision resistors fall within this category and are usually considered to be resistors which are manufactured to tolerances of ± 1 per cent or less and which are stable during their normal life and operating conditions to within toler-

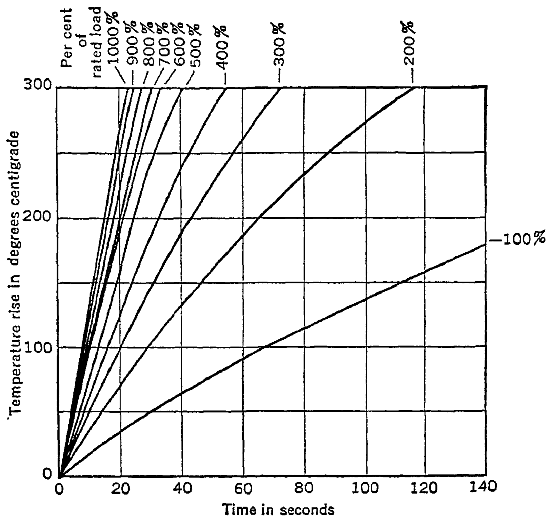


Fig. 8. Heating Time Required to Raise Hot-spot Temperature 300 Deg Cent for 50-watt (9/16 in. x 4 in.) Resistor for Overloads up to 1000 Per Cent of Rated Load. A 100 per cent load will bring full 300 deg cent rise in about 10 minutes. (Courtesy Ohmite Mfg. Co.)

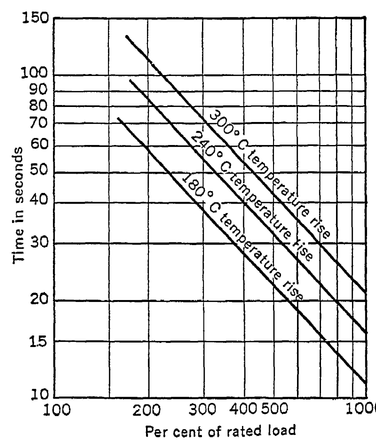


Fig. 9. Heating Time Required to Raise Hot-spot Temperature 180, 240, and 300 Deg Cent for Overloads up to 1000 Per Cent of Rated Load. Based on rating of 50 watts for 9/16 in. x 4 in. vitreous enamel resistor. Will reach 300 deg cent with continuous operation at rated load. (Courtesy Ohmite Mfg. Co.)

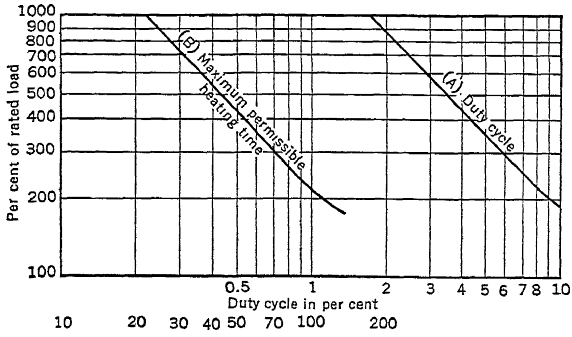


Fig. 10. Permissible Duty-cycle for 50-watt (9/16 in. x 4 in.) Resistor Based on a Maximum Hot-spot Temperature Rise of 300 Deg Cent. Curve A—Per cent of rated load at which resistor may be operated on intermittent duty. Curve B—Maximum time resistor may operate during any one cycle. (Courtesy Ohmite Mfg. Co.)

ances of something less than the initial adjustment. A few of the types in general use today are the following:

Spool Type, in which the winding is applied to a core body in the form of a spool, either molded or fabricated, to the dimensions necessary to accommodate sufficient wire for maximum resistance value desired and to dissipate rated watts with the minimum of self-heating. The spool may be divided into sections to control the voltage gradient between individual turns and between sections of winding and to permit some improvement in frequency characteristic over a winding applied in a single section. Typical sizes available are shown in the tabulation.

Spool Dimensions, in.		Watts Rating	Maximum Resistance Value, megohms Minimum 0.0015 in.-diameter Wire
Diameter	Length		
19/32	19/32	0.5	0.15
19/32	1 1/32	0.75	0.4
25/32	1 9/32	1.0	1.0

Flat-type resistors (card-type) usually consist of a single layer of bare or insulated wire, wound on a flat card of insulating material. The wound card is either encased in molding compound or is given a protective coating of lacquer or varnish. Maximum resistance value is typically of the order of 5000 ohms per inch of length and wattage dissipation about 1 watt per square inch of radiating surface.

Pad-type resistors consist of multiple windings on a spool or card to form resistor networks suitable for use as attenuators.

Plug-in type resistors are provided with terminations at one end in the form of a vacuum-tube base, providing means for inserting attenuator pads in a circuit.

Flexible resistors are made in the form of a helix wound on a flexible insulating cord which in turn is insulated with a flexible sleeving.

WINDING TYPES. Numerous types of winding are commonly used in resistors in an effort to attain low residuals of capacitance and inductance or simply to achieve the desired resistance value in the most economical way. Some of these winding types are shown here. Typical residuals obtained are shown in Table 2.

Table 2. Net Residuals of Typical Resistor Windings

Winding Type	Resistance, ohms	Approximate Net Residual
Continuous (flat-card type)	0–4.5	0–1.5 μh
	4.5–20	0.5–3.5 μh
	20–600	3.5–15 μh
	1,000	70 μh
	10,000	±3 μμf
Continuous (tubular power type)	0– 200	2–15 μh
	200– 800	15–60 μh
	800–2,000	60–130 μh
	2,000–5,000	130–500 μh
Bifilar	10	0.5 μh
	100	1.5 μh
	1,000	100 μμf
	3,500	300 μμf
Reverse layer	1,000	10 μμf
	3,500	30 μμf
	10,000	100 μμf
Reverse section	1,000	100 μh
	3,500	300 μh
	10,000	±5 μμf
Reverse half-section	1,000	±5 μμf
	3,500	±5 μμf
	10,000	30 μμf
Mandrelated filament	1,000	1–2 μμf
	3,500	1–2 μμf
	10,000	1–3 μμf
	35,000	2–4 μμf
	100,000	2–5 μμf
	200,000	3–8 μμf
Parallel opposing (Ayrton-Perry)	100	0–1 μh
	1,000	1–2 μh
	10,000	1–2 μμf

Continuous Winding. The simplest of all windings is one in which the wire is wound on a core continuously in one direction, known generally as a continuous or inductive winding. If the wire is wound in layers, the residuals of both capacitance and inductance are large and the resistor can be used merely for very low frequencies or direct-current applications. By winding in a single layer on a core of small cross-sectional area such as a flat card or small-diameter core, the inductance and capacitance are reduced. By spacing the turns and keeping the wire size to the minimum almost any desired residual may be achieved with the obvious limitation that short lengths of wire become increasingly difficult to adjust.

Bifilar Winding. This winding is used most generally where low inductance is required. In this winding the wire is bent back on itself at the midpoint so that the two half-lengths are side by side, separated by the insulation only. In this straight form any given wire has the minimum possible inductance. This minimum inductance L may be computed from the formula:

$$L = 0.005l(2.303 \log_{10} \frac{2D}{d} + 0.25)\ \mu\text{h}$$

where l = total length of wire in inches, D = distance apart, between centers, d = diameter of the bare wire, and μh = microhenries. D and d are expressed in the same units.

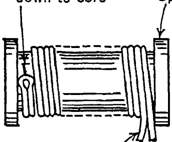
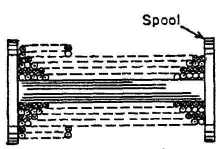


FIG. 11. Bifilar Winding.

In practice this loop is generally wound on a core (see Fig. 11), often in several layers, with the result that L is increased slightly. The bifilar winding has the disadvantage of high distributed capacitance. Next to a straight inductive winding, it is the cheapest winding to apply.

A modification of the straight bifilar winding is the series-bifilar winding in which the total length of wire is divided into sections, each wound separately as a bifilar loop and the sections connected in series. This has the advantage of maintaining minimum inductance and of reducing the capacitance by any desired amount, inasmuch as the capacitance is approximately inversely proportional to the square of the number of sections. This results in a good resistor, but it is expensive to apply, and since simpler windings generally give practically the same results it is seldom used in regular production.

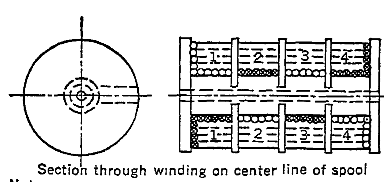
Reverse Layer Winding. The reverse layer winding (see Fig. 12) is used where the frequency range is low enough to permit appreciable residuals and where the spool is not divided into sections for a sectional winding. The direction of turns is reversed for each layer. It approximates the series-bifilar winding, but, because each layer must be secured in place before reversing turns, it is expensive where a large number of layers are applied. To be non-inductive it must have an even number of full as well as partial layers.



Section through winding on center line of spool

Notes:
1. Section shows alternate full layers and final partial layers wound in opposite directions. o indicates layer wound towards observer. ● indicates layer wound away from observer.
2. Winding always consists of even number of full and partial layers.

FIG. 12. Reverse Layer Winding



Section through winding on center line of spool

Notes:
1. ⊙ indicates section of winding wound toward the observer.
   ● indicates section of winding wound away from observer.
2. Approximately the same number of turns are wound in each of the sections.
3. Figures indicate winding order.

FIG. 13. Reverse Section Winding

Reverse Section Winding. The reverse section winding (see Fig. 13) is used largely for high resistance values. This winding is divided into two or more adjacent sections of equal size and number of turns. As in the series-bifilar winding, the distributed capacitance is approximately inversely proportional to the square of the number of sections. The turns in adjacent sections are wound in opposite directions to reduce the inductance, thereby approximating the reversed layers of a reverse layer winding. Generally the spool is sectionalized to simplify the winding. However, if the wire is large and the turns correspondingly few, the same effect is attained in a bunch-type winding in which the wire is wound in successive bunches around the core.

Reverse Half-section Winding. Although the reversed section winding gives excellent results for capacitance reduction, the magnetic coupling between adjacent sections is so low that in spite of reverse turns the inductance is often too high. This is especially true for values ranging from 1000 to 4000 ohms. Although above this upper limit high residual inductance still obtains, tolerable phase angle results because of the increasing effect of higher resistance value in the $L - CR^2$ relationship of eq. (6). Reversing by half-sections has the effect of increasing the capacitive and decreasing the inductive residuals. This winding is somewhat more expensive to apply than the reverse section type.

Parallel Opposing (Ayrton-Perry) Winding (See Fig. 14). This type of winding consists of two inductive windings on a core with turns equal and in opposite direction. A spaced layer of either insulated or bare wire is first applied in one direction. The second wire is wound in the opposite direction between the turns of the first winding. When bare wire is used, the cross-overs must be at exactly diametrically opposite sides of the core so that contact occurs at points of equal potential on each wire. The two windings are connected in parallel. The distributed capacitance is low, and the opposing currents in the two wires produce the minimum of magnetic effect. It has the disadvantage of requiring four times as much wire of any given size to obtain the same value produced from a normal single-wire winding.

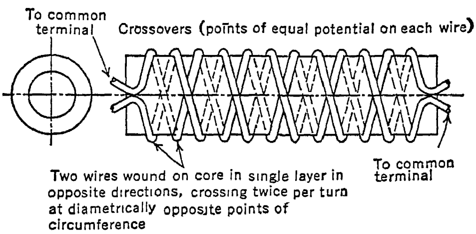


FIG. 14. Parallel Opposing (Ayrton-Perry) Winding

Mandrelated Filament Winding. This is an adaptation of the continuous winding discussed earlier, in which a single layer of resistance wire is wound on a flexible core. The helical filament so formed can then be handled in much the same manner as ordinary resistance wire. A number of such filaments have been developed with resistance wire, dimensions, and turn spacing so selected that for any straight length of filament the inductive component L/R is substantially equal to the capacitive component CR. The effect of winding on a form and terminating the filament is to increase the capacitance slightly. Because of the excellent frequency characteristics and the high resistance per linear length of filament (up to 2400 ohms per inch in 0.030-in. diameter of insulated filament), this type of winding is widely used for high resistance values in many of the designs developed for use in the communications field.

Miscellaneous Windings. Numerous other windings have been developed for use at high frequency that serve satisfactorily but are not in general commercial use except in precision-type measuring apparatus and in resistance standards. One of these is the "woven type" in which wire is woven into a cloth or ribbon pattern giving the effect of a continuous winding on a card of infinitesimal thickness. Rather intricate patterns are sometimes used to obtain low phase angle. Constructions such as "reversed turn" and "Curtis" windings require handling of each turn individually and therefore do not lend themselves to economic manufacture in mass production. Occasionally zero phase angle is achieved by adding the necessary amount of capacitance or inductance to a completely wound unit.

3. COMPOSITION CARBON RESISTORS

FIXED RESISTANCE RESISTORS. The designation "composition resistor" denotes a type of resistor that has very wide application in electronic equipment and apparatus because of its light weight, compactness, wide range of resistance values covered, and ease of mounting. It is primarily used in circuits where drift and variation in resistance value with time, temperature, humidity, and applied voltage are not of particular significance. In general, the resistive element in a composition resistor is a combination of finely divided carbon or graphite, a non-conducting inert material or filler such as talc, with synthetic resin as a binder. These substances are proportioned so as to yield the proper resistance value in the finished product.

Composition resistors are available in insulated and non-insulated types. In general, the large majority of composition resistors used in electronic equipment is of the insulated type. In this type the resistive element of the unit is surrounded by a substantial housing of insulating material, such as mineral-filled Bakelite, so that there is no possibility of contact with the element other than through the wire terminal leads of the resistor. The usual form of insulated resistors for wattage dissipations of 2 watts and less is a cylindrical body provided with axial terminals. In the uninsulated resistor, a cylindrical rod of resistive material is generally equipped with radial leads and the unit is painted.

SIZES AND RATINGS. Typical sizes in general usage of fixed insulated axial lead resistors with cylindrical bodies are as listed in Table 3.

Table 3. Sizes of Cylindrical Insulated Resistors

Nominal Wattage Rating at 40° C	Maximum Body Length, in.	Maximum Body Diameter, in.	Lead Length, in.	Min. Lead Diameter, in.
0.25	0.438	0.125	$1\frac{1}{2} \pm \frac{1}{8}$	0.028
0.50	0.438	0.156	$1\frac{1}{2} \pm \frac{1}{8}$	0.028
1.0	0.750	0.280	$1\frac{1}{2} \pm \frac{1}{8}$	0.032
2.0	0.750	0.344	$1\frac{1}{2} \pm \frac{1}{8}$	0.036

The resistors listed in Table 3 are available in resistance values ranging from a few ohms to 22 megohms on a manufacturer's standard basis, with the qualification that the 0.25-watt unit may not be obtainable in values as low as for the other units. These resistors are furnished in accordance with the RMA preferred number system for the values and tolerances shown in Table 4.

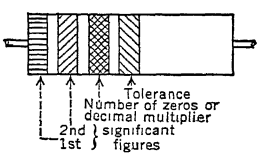


FIG. 15. Standard Color Coding

The resistance value is indicated by a color code applied to the resistor as shown in Figs. 15 and 16 and Table 5. The exterior body color of insulated resistors may be any color except black. It is recommended that the maximum continuous working voltage, either d-c or rms a-c, should not exceed 250 volts for the 0.25-watt unit; 350 volts for the 0.5-watt unit; and 500 volts for the 1- and 2-watt units.

Table 4. Nominal Resistance Values

(RMA preferred number system)

Nominal Resistance					Available in Tolerances ± per cent
ohms				megohms	
10	100	1,000	10,000	0.10, 1.0, 10	5, 10, 20
11	110	1,100	11,000	0.11, 1.1, 11	5
12	120	1,200	12,000	0.12, 1.2, 12	5, 10
13	130	1,300	13,000	0.13, 1.3, 13	5
15	150	1,500	15,000	0.15, 1.5, 15	5, 10, 20
16	160	1,600	16,000	0.16, 1.6, 16	5
18	180	1,800	18,000	0.18, 1.8, 18	5, 10
20	200	2,000	20,000	0.20, 2.0, 20	5
22	220	2,200	22,000	0.22, 2.2, 22	5, 10, 20
24	240	2,400	24,000	0.24, 2.4	5
27	270	2,700	27,000	0.27, 2.7	5, 10
30	300	3,000	30,000	0.30, 3.0	5
33	330	3,300	33,000	0.33, 3.3	5, 10, 20
36	360	3,600	36,000	0.36, 3.6	5
39	390	3,900	39,000	0.39, 3.9	5, 10
43	430	4,300	43,000	0.43, 4.3	5
47	470	4,700	47,000	0.47, 4.7	5, 10, 20
51	510	5,100	51,000	0.51, 5.1	5
56	560	5,600	56,000	0.56, 5.6	5, 10
62	620	6,200	62,000	0.62, 6.2	5
68	680	6,800	68,000	0.68, 6.8	5, 10, 20
75	750	7,500	75,000	0.75, 7.5	5
82	820	8,200	82,000	0.82, 8.2	5, 10
91	910	9,100	91,000	0.91, 9.1	5

Table 5. Color Code

Color	Figure or Number of Zeros	Decimal Multiplier	Tolerance, per cent
Black	0		
Brown	1		
Red	2		
Orange	3		
Yellow	4		
Green	5		
Blue	6		
Violet	7		
Gray	8		
White	9		
Gold		0.10	± 5
Silver		0.01	±10
No color			±20

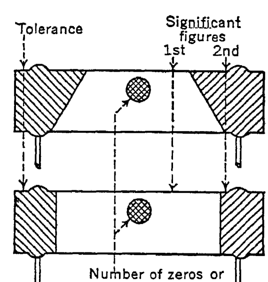


FIG. 16. Alternate Color Coding for Radial-lead Resistors

PERFORMANCE CHARACTERISTICS. In addition to obvious requirements on the mechanical properties of a resistor, such as ruggedness, security of terminals, legibility of color code, and its d-c resistance value and tolerance as established by a suitable d-c resistance measurement, the following properties are of interest for commercial uses.

Resistance-temperature Characteristic. Table 6 indicates the range within which an insulated composition resistor may be expected to vary at the temperatures indicated.

Table 6. Insulated Composition Resistor Variations Due to Temperature

Nominal Resistance Value	Maximum Per Cent Change in Resistance from Value at Ambient Temperature of +25 ° C			
	To −15° C Ambient	To −55° C Ambient	To +65° C Ambient	To +105° C Ambient
10 ohms to 1000 ohms	−0, +3.5	−0, + 6.5	±3	±5
1010 ohms to 10,000 ohms	−0, +5	−0, +10	±4.5	±8.5
10,100 ohms to 0.1 megohm	−0, +6.5	−0, +13	±4.5	±8.5
Over 0.1 megohm	−0, +10	−0, +20	±5	±10

Voltage Coefficient. This coefficient relates to the percentage change in resistance value per unit change in voltage with applied voltage as distinguished from any effects caused by heating at the applied voltage. It arises from a change in the conducting properties of the resistive material as the applied voltage is varied. Voltage coefficient is usually determined for resistors of 1000 ohms and above as follows:

$$\text{Voltage coefficient (per cent)} = 100\,\frac{R_1 - R_2}{R_2} \times \frac{1}{E_1 - E_2}$$

where E_1 = rated continuous working voltage, E_2 = 0.1 rated continuous working voltage, R_1 = resistance at rated continuous working voltage, and R_2 = resistance at 0.1 continuous working voltage. For resistors rated at $^1/_4$ and $^1/_2$ watt, the voltage coefficient should not exceed 0.035 per cent per volt, and for higher-wattage resistors it should not exceed 0.02 per cent per volt.

Humidity Effects. Test data indicate that molded-housing-type insulated resistors which have been thoroughly dried and then subjected to a condition of 90 per cent relative humidity at an ambient of 30 deg cent for 200 hours may in general be expected to stay within a limit of about 5 per cent, the magnitude of the change depending upon resistance value. The product of certain manufacturers is also capable of withstanding exposure to 100 per cent relative humidity at an ambient temperature of 66 deg cent for 250 hours without changing in resistance value by more than 10 per cent.

Noise. Current noise arises within the resistive element of a composition resistor primarily because of the microphonic nature of particle-to-particle conduction of current in the structure. In some resistors noise may also originate at the junction between the lead wire and resistive element because of imperfect contact.

It has been found that the product of certain manufacturers will have root-mean-square values of current noise with rated d-c voltage applied to the resistor terminals less than 3 rms microvolts per volt for $^1/_2$-watt units up to a resistance value of 1 megohm, and 6 rms microvolts per volt for resistance values above 1 megohm. For 1- and 2-watt resistors, the noise level may be expected not to exceed 1.2 rms microvolts per volt.

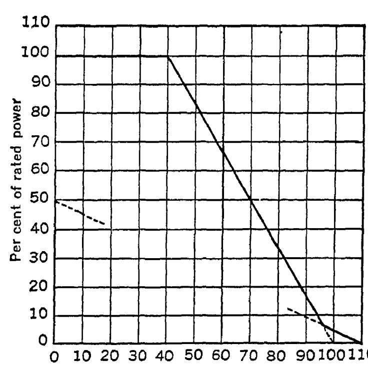


FIG. 17. Derating Curve for High Ambient Temperatures

Load-life Characteristic. D-c load tests conducted at 40 deg cent have shown such extremely wide variations in resistance change in 1000 hours that no general statement is of value. Units have been found to age either positively or negatively, the film-type resistor usually showing an increase in value and the body type a decrease in value. Hence the manufacturer should be consulted for specific information on the performance of his product. However, resistors are available which will not change more than 10 per cent under this 1000-hour load test.

For resistors designed to carry 100 per cent of rated load at 40 deg cent ambient, it is recommended that the derating curve shown in Fig. 17 be followed to prevent undue aging.

It will be noted, for example, that resistors functioning in an ambient of 85 deg cent should be operated at not more than 25 per cent of rated load. Resistors are also available which may be operated at full wattage at 70 deg cent with a derating to zero wattage at 150 deg cent.

Effect of Soldering. Care must always be taken to prevent the quality of a composition resistor from being seriously impaired and its resistance value and stability affected significantly by excessive heating during the operation of soldering its terminal leads to apparatus or equipment terminals. The leads should be left as long as possible and preferably not less than $^3/_8$ in.

Additional Characteristics. In addition to the properties and tests described above, other characteristics of interest in military work and of possible importance in certain commercial application may be mentioned: voltage breakdown strength of the insulation on insulated-type resistors; high-altitude flashover voltage; performance after salt-water-immersion cycling; effects of thermal shock under temperature cycling in air; effect of mechanical vibration on mounted resistors; performance under other types of load-life tests; short-time overload performance; and ability to withstand specific mechanical tests on the security of terminals. Further details on tests and expected performance may be found in Specification JAN-R-11.

FREQUENCY CHARACTERISTIC. In general the inductance associated with composition resistors is sufficiently small to be disregarded. At low frequencies the value of resistance R is the same as the d-c resistance, but as the operating frequency is raised the value of R starts to decrease and may reach a value which is only a few per cent of the d-c value. The frequency at which R begins to show a significant decrease in value depends upon its d-c value; that is, the greater the d-c value of a resistor, the lower the frequency at which departure from the d-c value is observed. Incidentally, parallel capacitance C shows a similar decline in value, but the amount of change is significantly less than the reduction in resistance.

Figure 18 shows the variation in resistance with frequency of two different 1-megohm insulated resistors of $^1/_2$-watt rating. These resistors are practically identical in physical size, but one has a resistive element in the form of a composition film on the surface of a glass tube and the other has a resistive element of the body type. Resistors that have been found to show the least decrease with frequency have a very thin uniform resistive film on a rod or tube of low-loss dielectric material in which the ratio of length of element to its diameter is large and in which no exterior covering or coating material is in contact with the surface of the element.

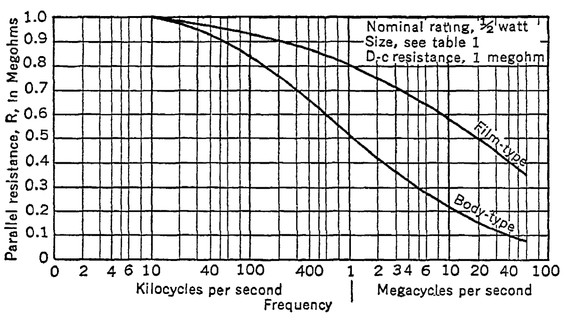


FIG. 18. Variation of Parallel Resistance R with Frequency for 1-megohm Molded Insulated Housing Type of Resistor

It has been observed that, in a given specific type of resistor, the ratio of parallel a-c resistance to the d-c resistance is approximately constant for a fixed value of the product of operating frequency and d-c resistance value. That is to say the ratio is approximately the same for a 10-megohm unit operating at 0.1 mc, a 1-megohm unit at 1 mc, and a 0.1-megohm unit at 10 mc. In each instance the product of megohms and megacycles is unity. Applying this example to the curve given for the body-type resistor in Fig. 18, and noting that at 1 mc the 1-megohm resistor has 53 per cent of its d-c resistance, a resistor having a value of 10 megohms (d-c) and being of the same construction as the 1-megohm unit for which the curve is drawn would have a parallel resistance of around 5.3 megohms at 0.1 mc. Also a 0.1-megohm unit (d-c) at 10 mc would have a value of about 53,000 ohms parallel resistance. It is very important to observe that this relationship is only approximate and also must be established for each specific type of resistor. This is evident from the fact that, for a product of unity for the film type of resistor shown, the parallel resistance is about 80 per cent of the d-c value, as compared to 53 per cent for the body-type unit.

Table 7 indicates the approximate frequency at which a 1-megohm resistor will have a parallel resistance of 0.5 megohm for 0.5-, 1- and 2-watt resistors within the dimensional limits stated in Table 3. With the product relationship discussed above, the data in Table 7 may be used to estimate the frequencies at which other resistors of the same size and type will drop to values of parallel resistance equal to half of the d-c resistance.

The parallel capacitance C associated with these resistors is also a function of resistor construction, resistance value, and frequency. Considerable variation in capacitance value for different samples of the same type has been noted, but the value appears to be in the order of 1 to 3 $\mu\mu f$, falling with frequency to around 0.5 to 1 $\mu\mu f$.

HIGH-FREQUENCY AND HIGH-VOLTAGE TYPES. Another line of composition-type resistors is available which is designed for high-voltage and high-frequency applications. These have various wattage ratings from less than 1 watt to about 100 watts.

In general the high-frequency type consists of a continuous film of resistive element on the surface of a ceramic tube. In size, as typified by the product of one manufacturer, these resistors range from units about $1/8$ in. in diameter and $1/2$ in. long with wire lead axial terminals to units about 2 in. in diameter and 20 in. long overall, including ferrule terminals. Resistance values of a few ohms to several megohms are available, depending upon the physical size of the resistor.

Table 7. Resistance-frequency Characteristics of Specific Resistors

1-megohm Resistor Nominal Wattage Rating	Frequency in Megacycles for R of 0.5 Megohm	
	Film-type Element	Body-type Element
0.5	20	1.2
1.0	3.0	0.8
2.0	....	0.4

High-voltage composition-type resistors are of similar construction to that described for the high-frequency type, except that the resistive element is usually in the form of a spiral or ribbon to provide a long conducting path. They are available in sizes ranging from units $5/16$ in. in diameter and about 2 in. long rated at 2 watts and 7500 volts maximum, to units 2 in. in diameter and about 20 in. long rated at 150 watts and 100,000 volts maximum. Resistance values obtainable range from a few thousand ohms to a million megohms.

4. DEPOSITED-CARBON RESISTORS

GENERAL DESCRIPTION. "Deposited-carbon resistor" denotes a kind of resistor in which the resistive element is a film of carbon deposited on the surface of a suitable ceramic core by the thermal decomposition of gaseous hydrocarbons at high temperatures. This film is extremely thin and by proper control of the coating process may be varied in thickness within the range from 1×10^{-4} to 5×10^{-8} in. The resistance of such films ranges from about 5 ohms per unit square to about 10,000 ohms per unit square. "Ohms per unit square" denotes the resistance as measured between opposite edges of a square of resistance film of the thickness indicated.

Electrical connection is made to the carbon film by applying low-resistance electrodes of either graphitic or special metallic paint. These electrodes are cured by suitable heat treatment, and then metal caps with integral lead wires are forced over them. Since the carbon film is sensitive to abrasion and also needs protection from contamination, the resistance structure is coated with a suitable baking varnish or housed in an enclosure. Resistors having resistance values up to a few thousand ohms are formed of uniform films. High-ohmage units are obtained by cutting a helical groove through the carbon film to form a ribbon of film wound around the core between the end electrodes. The surface perfection of the ceramic core and its other physical and chemical properties have a marked effect on the electrical characteristics of the resistor.

USAGE. Deposited-carbon resistors provide exceptional resistance stability and compactness in high values of resistance. Also, because of low residual inductance, the power-type varieties are of value as load resistors in testing high-frequency equipment. Special shapes of deposited-carbon resistors in the form of suitably terminated small rods and disks are available for assembly in coaxial-type attenuator units for use in making accurate attenuation measurements up to frequencies of several hundred megacycles.

SIZES AND RATINGS. Typical figures for the product of one manufacturer are given in Table 8. These resistors have cylindrical bodies. The resistors with the shell enclosures have axial leads and are intended for general-purpose use. The glass-enclosed units are hermetically sealed with an inert gas in the enclosure and are provided with ferrule end caps as terminals; they are intended for applications in which high stability in resistance value or high levels of power dissipation are involved.

PERFORMANCE CHARACTERISTICS. The general-purpose resistors and the glass-enclosed types for high-stability application are available in tolerances as close as ±1 per cent. The large power dissipating units are available in tolerances of ±5 per cent.

Table 8. Ratings and Sizes of Deposited-carbon Resistors

Rating Nom. at 30° C	Wattage, Maximum *	Approximate Overall Size, in.		Resistance Range, ohms		Short-period Peak Voltage †	Protective Enclosure
		Diameter	Length	Minimum	Maximum		
0.15	0.5	11/64	13/16	1	5×10^6	5,000	Shell
0.5	1.5	11/32	1	200	10^6	8,000	Shell
1.0	3.0	11/32	2 1/16	200	5×10^7	15,000	Shell
1.0	10	7/16	2 1/4	200	10^7	2,000	Glass
2.0	20	7/16	3 1/4	200	1.5×10^7	6,000	Glass
	60	5/8	4 11/16	20	10^7	10,000	Glass
	300	1 1/4	8 3/4	20	5×10^6	20,000	Glass
	600	1 1/4	14 3/4	40	10^7	40,000	Glass

* At maximum wattage, resistance may differ from 30 deg cent value by decreases of 10 to 15 per cent.
† Not to exceed that required for maximum power rating.

Temperature Coefficient. The temperature coefficient of the carbon film depends upon its thickness and ranges from about −180 parts per million per degree centigrade for very heavy coatings to about −500 parts per million per degree centigrade for light coatings. Furthermore, the application of protective lacquer to the film may modify its temperature coefficient by virtue of mechanical effects on the film. The physical properties of the ceramic base also affect this coefficient. Hence the temperature coefficients of these resistors range from about −0.02 to possibly −0.10 per cent per degree centigrade, depending upon resistance value and constructional features. The resistance-temperature curve is approximately linear over the temperature range of −40 deg cent to +60 deg cent.

Voltage coefficient is in general negligible for deposited-carbon units. Occasionally an individual resistor may show a slight resistance variation with voltage but probably not more than 0.002 per cent per volt.

Humidity effects depend upon the structure of the resistor and in hermetically sealed units become a matter of leakage across the surface of the housing. The following figures are indicative for general-purpose units with shell enclosures. After exposure to a condition of 90 per cent relative humidity at an ambient of 30 deg cent for 200 hours, the maximum change in resistance may be expected to be less than 1 per cent, with an average change of less than 0.5 per cent.

Noise. At low levels of voltage, deposited-carbon resistors exhibit pure thermal noise, but as the voltage is raised other electrical noise may be observed. This may be some form of contact noise due to imperfections or loose contacts in the carbon film. However, at rated load the noise level, excluding thermal noise, may be expected not to exceed 0.25 rms microvolt per volt.

Load-life Characteristics. General-purpose-type deposited-carbon units when operated at their normal wattage rating may be expected to show average changes in resistance value of not more than 1 per cent after 2000 hours. Occasional units may age as much as 1.5 per cent. Hermetically sealed resistors operated at high power levels may show more rapid aging. For example, the 7/16 × 3 1/4 unit in Table 8 may change in resistance value up to 3 per cent after 3000 hours of operation at 10 watts. Load aging may cause either a positive or negative change in resistance but generally results in a positive change. In using power-type resistors, care must be taken to insure that the voltage gradient within the resistor is not sufficient to give rise to corona or cause flashover between adjacent turns of a spiraled element and thus damage the unit.

Like composition resistors, when deposited-carbon units are operated in high ambient temperatures they should be derated. For types which are not hermetically sealed, it is recommended that the maximum operating surface temperature of the carbon film not exceed 120 deg cent. Since their normal rating is established at 30 deg cent, the wattage ratings for these units should be decreased by about 1 per cent for each degree centigrade that the ambient exceeds 30 deg cent. For glass-sealed units, which under maximum power ratings may operate at surface temperatures up to 450 deg cent, the ratings can be regarded as independent of ambient temperature up to about 80 deg cent. The manufacturer should be consulted for further derating information for the specific conditions of application. The ratings of power units can be increased several fold by forced air cooling, and, for a-c application, resistors are available which may be liquid cooled through direct contact of the coolant with the film. Anodic oxidation of the film precludes the use of water cooling on d-c applications.

The effect of soldering is negligible on lead-type units, provided, of course, that the resistor is not damaged by direct contact with the soldering tool.

No-load Aging. Under conditions of no-load or shelf-aging, deposited-carbon resistors of the general-purpose type may be expected to drift in value not more than 0.1 to 0.2 per cent per year. Hermetically sealed units appear to have a stability in resistance value of the order of 0.005 per cent per year, which is about the limit of error in measurements extending over such a time period. In fact the stability of high-quality hermetically sealed carbon film resistors, particularly in the megohm region, appears to be at least as good as, if not better than, that of equivalent resistance wire-wound units.

FREQUENCY CHARACTERISTIC. Deposited-carbon resistors exhibit a decrease in parallel resistance with frequency. However, the rate of decrease does not appear to be as rapid as in the insulated-type composition resistor, and it is less for the glass-enclosed type than for the varnished or shell-protected unit in which the protecting material is in contact with or very close to the carbon film. A 1-watt varnish-coated 1-megohm unit, $9/32$ in. in diameter and $2\ 1/16$ in. long, drops to 0.5 megohm parallel resistance at about 12 mc.

Whereas the inductance of an unspiraled film is negligible, the inductance of a spiraled unit may be appreciable. For example, the inductance of a 1-megohm resistor of the $7/16$ in. $\times$ $3\ 1/4$ in. size in Table 8 is about 1.1 microhenries. Though spiraling increases inductance, the ratio of inductance to resistance remains essentially unchanged. Since the effects of distributed capacitance are changed only slightly by spiraling, the alteration in high-frequency behavior of a resistor which is spiraled to a high resistance value is largely that associated with the resistance increase alone.

5. METAL FILM RESISTORS

These resistors are formed by coating a base of suitable material with a very thin film of metal or metallic alloy. This film may be applied to the base by cathode sputtering or metal evaporation processes, and by chemical methods. It is quite thin, being of the order of 10^{-7} in. thick. The material used for the base may be ceramic, glass, or an organic body such as Bakelite.

In one construction, a thin film of palladium is deposited by a chemical process on a ceramic-core tube. The ends of the film are coated with silver or some other metal to form electrodes for the attachment by soldering of radial lead wires. The film is covered by a form of vitreous covering to afford protection against oxidation and corrosion as well as against mechanical damage. High values of resistance are obtained by cutting a helix through the metal film into the core, in the same fashion as deposited-carbon film resistors are helixed. For further protection the whole unit, including the portion of the lead wires in contact with the resistor body, is lacquered. Resistors made in this way are available over a range of resistance values in several sizes corresponding to different wattage ratings. They possess good electrical stability under aging and load conditions, and they have good heat-dissipating properties. As determined from observations on a limited number of samples with a nominal rating of 1 watt, aging under d-c rated load is predominately positive as regards resistance value and may be of the order of 1 to 5 per cent after 1000 hours, depending upon resistance value. The temperature coefficient is negative and varies to some extent with resistance value. It also shows considerable departure from linearity over the temperature range of -40 deg cent to $+80$ deg cent. Referred to resistance value at 20 deg cent, resistors increase in value by about 5.5 per cent at -40 deg cent, and at 80 deg cent they decrease in value by about 3 per cent.

In another construction of metal film resistor, a resistance alloy such as Nichrome is evaporated to form a very thin film on the surface of a glass tube or rod. A layer of protective material is placed over this film, and lead wires are attached in the same general manner as described above. An advantage of this type of construction is that the low temperature coefficient of the resistance alloy is retained in the resistor.

6. POTENTIOMETERS AND RHEOSTATS *

Potentiometers and rheostats in general use in the communications and electronics fields may be classified as wire-wound and composition types. Those most extensively used are the continuously adjustable type, where a movable contact traverses a resistance-wire winding or a composition resistance element in small increments of their lengths. In most instances the movable contact is controlled by a rotatable shaft, although in some the resistance element is rotated. It is standard practice to have terminals for both ends

* Article 6 was contributed by A. H. Volz.

of the resistance element in addition to a terminal for the movable contact so that they may be wired either as potentiometers or rheostats.

Most of the potentiometers and rheostats hereinafter described may be obtained in tandem arrangement either with a common shaft which adjusts each unit simultaneously or with concentric shafts which permit independent adjustment of the tandem units.

WIRE-WOUND POTENTIOMETERS AND RHEOSTATS. **Low-operating-temperature potentiometers** find application as voltage- and current-adjusting devices in low-energy circuits. The power ratings are based on operation in free still air at an ambient temperature of 40 deg cent with a maximum temperature rise of 60 deg cent. A power derating curve is shown in Fig. 19. When potentiometers are enclosed and in close proximity to other components it is considered good practice to limit the power dissipation to about one-half the rating.

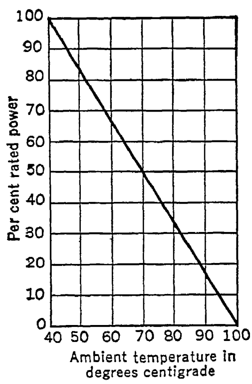


FIG. 19. Power Derating Curve for Continuous Duty

The types that find the widest application are the small circular potentiometers that range from about 1 1/4 in. to 2 in. in diameter and are rated from about 2 to 4 watts. In this type bare wire is space wound on a flat strip of insulating material, usually a laminated phenolic, which is then formed into a circular shape and set into a housing or case. The contact shoe or brush is made of a base metal or an alloy of base metals, supported on a spring member. The force exerted on the resistance wire by the base-metal contacts usually ranges from about 100 to 200 grams, sufficient to keep the contact resistance below 1 ohm, providing a satisfactory level of contact noise for most applications and a useful life of 25,000 to 100,000 cycles of operation. Individual potentiometer designs may show changes in resistance from 1 to 10 per cent over the ranges of temperature and humidity usually encountered.

In general it is desirable to use no smaller than 1.75 mil wire (usually similar to Nichrome), resulting in maximum resistance of about 50,000 ohms for the 2-in.-diameter size and 10,000 ohms for the 1 1/4-in. size. Standard tolerances on resistance are ±10 per cent. Closer tolerances and higher resistance values can be obtained on a custom basis. Finer wire is more susceptible to wear and electrolytic action.

Various resistance-rotation characteristics are available, the most extensively used being the linear type wherein the rate of change of resistance between the contact terminal and one of the end terminals is approximately constant with angular rotation. In these small types the degree of linearity is usually of the order of ±5 per cent of total resistance. This type of characteristic is illustrated by curve *A* of Fig. 20.

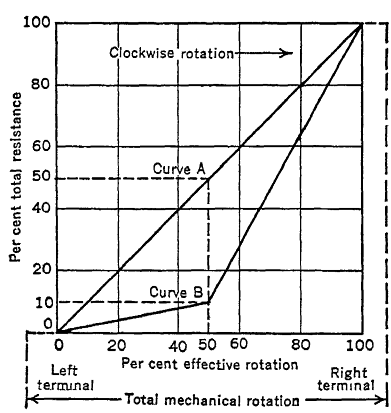


FIG. 20. Clockwise Tapers

For special applications, non-linear or tapered potentiometers can be obtained. This non-linear characteristic is effected by changing the pitch of the winding at certain points in the range of rotation or by changing wire size or by a combination of both. Curve *B* of Fig. 20 illustrates a clockwise and non-linear characteristic in which there are two sections differing in their rates of change of resistance. As a high percentage of the total resistance may be concentrated in a small portion of the resistance element, it is recommended that non-linear controls be rated at about half of the rated power of linear controls of the same designs.

For continuous operation of linear types, the maximum current through the entire resistance element or through any portion thereof should not exceed the value given by the following equation:

$$I = \sqrt{\frac{W}{R}}$$

where W is the wattage rating, R is the total resistance, and I is the maximum permissible current. The maximum current for each section of a non-linear control should be determined from the above expression except that R, instead of being the actual total resistance, should be that value which would obtain if the entire resistance element were wound with the same size wire as the particular section.

There is another group of circular, wire-wound, low-operating-temperature potentiometers which are approximately 3 in. in diameter and range from about 1 $^1/_4$ in. to 2 $^5/_8$ in. in depth, and in power rating from 8 to 15 watts. The winding is usually clamped around a molded phenolic base. This class of potentiometer is usually provided with a winding of a higher degree of uniformity than the smaller types, owing to the use of higher-precision winding methods. As a result linearly wound potentiometers are obtainable with linearity ranging from 1 to 0.3 per cent. Non-linear types are usually produced by shaping the card or strip on which the wire is wound, which is possible because of the greater depth of the winding. The amount of taper is limited by the maximum slope (angle between winding axis and shaped side of card) that can be wound without having the winding collapse which for close winding is about 40 deg. The deepest linear unit of this group can be wound to about 200,000 ohms without resorting to Nichrome wire finer than 1.75 mil in diameter. These types are readily obtainable with low-temperature-coefficient wire such as Advance, which, however, will wear more rapidly than Nichrome since it is softer. An idea of the performance capabilities of potentiometers of these two groups can be obtained by referring to the Joint Army-Navy Specification JAN-R-19.

Very special high-precision, low-operating-temperature wire-wound potentiometers have been developed for military applications. Some of them have winding cards shaped to provide special resistance-rotation characteristics to an accuracy represented by two turns of the winding at any point in the range of rotation. Some types have been provided with closed windings for continuous rotation; in one such type an input d-c voltage is applied through two fixed taps 180 deg apart and the output voltage is obtained from two rotating brushes diametrically opposed to each other. If the winding is linear, varying the position of the brushes varies the output voltage in accordance with a linear sawtooth wave.

Many of these types have been provided with precious-metal contacts developed specifically to obtain low contact resistance with low contact pressure. Contact resistance of a few hundredths of an ohm with contact pressures of about 50 to 80 grams has been obtained, and a life of the order of a million operations has been realized. One type of contact alloy widely used is Paliney No. 7, which consists of platinum, palladium, gold, silver, copper, and zinc.

In another type of low-operating-temperature wire-wound potentiometer, known as a multiturn potentiometer, the resistance wire is wound on an insulated metal mandrel about $^1/_8$ in. in diameter which is then formed into a helix. The diameter and number of turns of the helix vary, and potentiometers have been produced commercially with 2, 10, 15, 25, and 40 turns and with overall diameters of approximately 2, 3, 4, and 6 in. The contact is arranged to follow the path of the helix. With this type very fine adjustments are possible because of the comparatively long winding. By maintaining close tolerance on the diameter of the mandrel and on the resistivity and diameter of the wire, linearity of the order of ± 0.1 per cent or better can be provided.

The straight winding type potentiometers heretofore described are not very suitable for high-frequency applications. At frequencies above the audio range the straight winding controls are affected by distributed capacitance and inductance. For higher-frequency applications the characteristics of the individual control should be investigated.

POWER-TYPE RHEOSTATS. Toroidal Winding Type. This type is the most extensively used for applications from 25 to 1000 watts in the communications and electronics fields. Adjustable resistors of this type generally consist of a toroidal ceramic form which is wound with either round or ribbon-type wire over an arc of approximately 300 deg. The wound form is then placed in a suitable ceramic base, and the entire unit, except for the contacting surface of the wire, is given a coating of vitreous enamel under high temperature. This form of construction permits a high wattage rating in a relatively small volume of space.

Rheostats of this type are available in sizes ranging from approximately 1 $^1/_2$ in. to 12 in. in diameter and in rated wattage from 25 to 1000 watts. They are wound to resistance values from a fraction of an ohm up to 10,000 ohms. The power rating of rheostats is based on temperature rise in free still air. For those rated at 100 watts or less, the temperature rise is limited to 300 deg cent; for those rated above 100 watts the permissible temperature rise is 350 deg cent.

These rheostats are available with either linear or tapered windings. Specific uses for tapered rheostats are: (1) to provide a more uniform degree of control for all positions of the contact, (2) to make possible the use of a smaller control, (3) to make it possible to wind a higher resistance on a small control for specific applications, (4) to provide a particular controlled effect. An example of this last might be to provide a linear relationship between control setting and motor speed in the case of motor speed controls, or to give linear control of light output from a lamp. Figure 21 shows how the current varies (in three typical rheostats) with per cent rotation of the contact.

For special applications, controls can be obtained with continuous 360-deg windings, built-in toggle switches, or off positions at either end of the rotation.

Metal Type. Another power type of rheostat utilizes mostly metal in its construction. The wire or ribbon is wound on a strip of aluminum with asbestos as insulation between the wire and aluminum strip. The winding is formed into a circular shape and is assembled in a die-cast aluminum base with mica separating the winding from the base. Owing to the close proximity of the winding to the aluminum parts, the heat is carried away from the winding more rapidly than in the ceramic types. As a result, for the same wattage dissipation the temperature rise is somewhat lower.

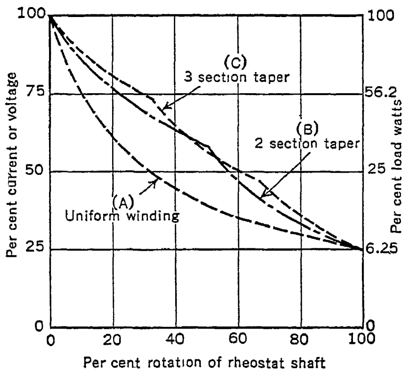


FIG. 21. Typical Curves of Current, Voltage, or Wattage Relative to Shaft Rotation for Uniformly Wound and Taper-wound Rheostats (Courtesy Ohmite Mfg. Co.)

Standardization requirements and performance capabilities of power-type rheostats for use in military applications are contained in Joint Army-Navy Specification JAN-R-22.

Tubular Slide-wire Type. The tubular slide-wire type of rheostat is used extensively for general laboratory applications, particularly for precision measurements. They are not generally used in commercial applications, inasmuch as they require considerably more mounting space for equivalent wattage ratings and are not as convenient as the toroidal type.

STEP-TYPE POTENTIOMETERS, RESISTANCE BOXES, AND ATTENUATORS. **Potentiometers.** This type consists of a rotary-type switch wired with fixed wire-wound resistors between successive contact positions. A brush or blade rotated by a shaft makes contact with contact studs or clips to which the fixed resistances are wired. The resistors of each step may be all of equal resistance value or they may differ, depending upon the nature of the application. The use of finite resistance steps permits a high degree of accuracy, especially in low- and medium-frequency applications. The accuracy at higher frequencies is dependent upon the frequency characteristics of the individual resistors and the type of switch structure. Likewise, the amount of power that can be dissipated is governed by the fixed resistors and the type of switch used.

Decade Resistance Boxes. Many laboratory measurements and test instruments require the adjustment of resistance in a circuit in accurately known steps of pure resistance. A decade resistance box consists of a number of individual resistance units equipped with switching so that the total resistance is adjustable in decade units. Resistance units are connected in series, and as many units can be connected as are required. To keep minimum inductance and distributed capacitance, card-wound or spool-type resistor units are used. Shielding and careful wiring arrangement also help keep capacitance low between resistor units. Where residual inductance of the resistors must be considered, a switching arrangement introduces a compensating winding as the resistance is adjusted to maintain constant inductance.

Attenuators. Attenuators are used to insert known amounts of transmission loss in circuits either for testing purposes or for volume level control. Step-type attenuators basically consist of rotary-type switches and fixed resistors, as in step-type potentiometers. They are arranged, however, to introduce various types of balanced or unbalanced resistive networks into a circuit. They are designed electrically to be inserted between specific input and output impedances, and only when so used will they insert the desired loss (see Section 5).

CARBON COMPOSITION TYPE POTENTIOMETERS. Carbon composition potentiometers are widely used in the communications and electronics fields on account of their low cost, the higher resistance values in which they can be obtained, and their excellent high-frequency characteristics. The types generally available are physically similar to the small single-hole mounting, low-operating-temperature wire-wound controls previously described. Two types of composition resistance elements are used, namely, the film-coated type and the molded type. In the film-coated type, the carbon, filler, and binder mixture are applied as a film on a ring of insulating material. The film is specially processed so as to minimize abrasion of the contact surface of the resistance element. In the molded type the carbon composition is molded into a phenolic base. The contact is a carbon brush, giving a carbon to carbon contact.

Linear and non-linear resistance rotation characteristics are obtainable in the composition types. The non-linear or tapered characteristic is produced by varying the proportion of the conducting material to the insulating material in the mixture as the element progresses over its length. It is possible by blending in this manner to obtain a rather smooth rate of change of resistance with angular rotation. Typical resistance-rotation characteristics of composition-type potentiometers are shown in Fig. 22. Curve *A* represents a clockwise linear characteristic except for a small range at each end of the rotation. Curve *B* illustrates a clockwise taper in which the first 50 per cent of the rotation introduces only 10 per cent of the resistance into the circuit, whereas the second half of the rotation inserts the remaining 90 per cent of the resistance. Curve *C* represents a clockwise taper in which the first 50 per cent of the rotation introduces 90 per cent of the resistance and the second half of the rotation inserts the remaining 10 per cent of the resistance. Curves *D*, *E*, and *F* illustrate counterclockwise tapers of similar characteristics to curves *A*, *B*, and *C*, respectively.

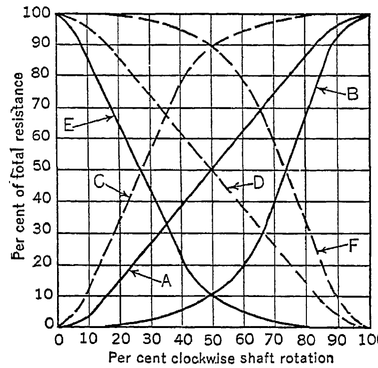


FIG. 22. Nominal Resistance-rotation Characteristics (Courtesy Ohmite Mfg. Co.)

Film-type composition potentiometers are available in a variety of sizes ranging from about 5/8 to 1 1/2 in. in diameter and in wattage ratings from about 0.05 watt to 1 watt. The available molded types are about 1 1/8 in. in diameter and are rated at about 2 watts. Resistance values obtainable range from about 50 ohms to 10 megohms.

For film and molded composition-type potentiometers, the voltage coefficient and the effects of overloading, aging, temperature changes, and exposure to high humidity are about of the same order as for film and molded fixed composition resistors. The molded are inherently more stable in resistance than the film types. The tabulation below obtained by tests compares typical film and molded types of 1-megohm resistance with respect to their stability of resistance under varying atmospheric conditions. The tabulation is in terms of average percentage change in resistance from that measured at room conditions in successive tests on the same set of samples.

Per Cent Change in Resistance from Initial Resistance at 20° C under Varying Atmospheric Conditions

	After 4 Hr at −18° C	After 4 Hr at −50° C	After 72 Hr at 50° C	After 72 Hr at 65° C	After 96 Hr at 40° C and 95% relative humidity	After 72 hr at 90° C
Film type	+3.0	+7.0	−7.0	−10.0	+14.0	−17.5
Molded type	+3.5	+5.5	−4.0	−5.5	+6.5	−5.5

Owing to their low inductance and capacitance, composition potentiometers are finding wide use in high-frequency applications; an example is a common shaft tandem arrangement of three rheostats, two of which are the series arms and the third the shunt arm of a continuously adjustable T type attenuator for level control in television transmission over telephone circuits at frequencies up to 4 megacycles.

7. SPECIAL-PURPOSE RESISTORS

In addition to the many resistors described so far, numerous special-purpose resistors designed for particular applications should at least be mentioned here. In this category are resistor standards such as those used in various measuring circuits and bridges which are discussed in Electrical Measurements, Section 11 of this handbook. Thermistors and varistors are used in an increasing number of applications in communications circuits, and they are fully described in the following articles. Resistors used primarily for heating as in furnace elements and heavy-duty power controllers are covered in the Electric Power volume of the *Electrical Engineers' Handbook*. A few other kinds are:

Resistance Lamps. In this type a special lamp filament having a high positive temperature coefficient serves either as a current-limiting, current-regulating, or protective device in communication circuits. Those used to maintain constant current, as in the heater circuits of vacuum tubes, are referred to as ballast lamps.

Dummy Antenna Loads. These may be in the form of (*a*) special space-wound resistance elements enclosed in an evacuated glass bulb, (*b*) a wound mica card mounted between metal castings to assist in carrying off the heat, or (*c*) a water-cooled film-type resistor in which water is circulated through or over a ceramic core on which a resistor film has been deposited.

RESISTORS IN PRINTED CIRCUITS. Printed circuit structures are a comparatively recent development in which a network of resistors, in combination with capacitors and inductors, is applied in the form of bands or ribbon to one or both sides of a supporting structure, usually a thin ceramic plate. The necessary capacitors are also mounted on the plate, and the plate itself may serve as the dielectric spacer and support for the capacitor. Connections to both resistors and capacitors are made by lines of conducting paint. Thus a complete coupling or filter circuit or even an entire amplifier circuit having several stages of amplification may be assembled on a thin plate of 1 or 2 sq. in. of surface area. The network is suitably terminated with wire leads, and the whole structure is given a suitable protective covering. The resistors may be of metallic or carbon film or of the composition film type. For composition and metal film type resistors produced by chemical means, the material may be applied in the desired pattern by painting or silk-screen process. For resistors applied by the metal-evaporation process, suitable stencils may be used to limit deposition to the desired areas. Where a high dielectric body serves both as the dielectric for the capacitors in the circuit and as the supporting panel for the resistors, note that the increased distributed capacitance brought about in the resistor due to the intimate contact with the high dielectric material may have a very marked effect on its performance at high frequency. Printed circuits are finding increasing applications in devices such as hearing aids, miniature radios, or wherever space is at a premium. Obvious advantages are compactness and small size. Disadvantages are difficulties in manufacture due to the necessity for processing all elements of the circuit at one time, which makes close tolerance adjustment of individual elements impracticable and lessens the likelihood of attaining the desired characteristics in all components. Also, in service, it is usually necessary to replace the entire unit when a single element becomes defective.

BIBLIOGRAPHY

Industrial Control Standards, Publication 1C4-22, July, 1946, National Electrical Manufacturers Assoc., 155 E. 44th St., New York, N. Y.
Blackburn, J. F., *Components Handbook*, Vol. 17, 1949, Radiation Laboratory Series, Massachusetts Institute of Technology, Cambridge, Mass.
Curtiss, H. L., and Grover, F. W. Resistance Coils for A.C. Work, *Bulletin Bureau of Standards*, Vol. 8, pp. 495–517 (1913).
B. Hague, *Alternating Current Bridge Methods*, 3d Edition, Chapters 2, 3, Pitman & Sons, 1932.
Bureau of Standards Circular 74.
Bureau of Standards Circular 100, 2d Edition, pp. 100–102.
Wireless Engineer, Vol. XII, No. 141 (June, 1935); No. 143 (August 1935).
High-Frequency Characteristics of Resistors, Report 520, Radiation Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.
Nettleton, L. A., and Dole, Fred E., Potentiometers, *Review of Scientific Instruments*, Vol. 17, No. 10, pp. 356–363 (Oct. 19, 1946).
Brunetti and Curtis, Printed Circuit Techniques, *Bureau of Standards Circular* 468.
New Advances in Printed Circuits, Bureau of Standards, Miscellaneous Publication 192.
Data Sources:
Driver-Harris Co., *Catalog* R-46.
Wilbur B. Driver Co., *Resistance Handbook.*
C. O. Jelliff Mfg. Co., *Resistance Alloys.*
Hoskins Mfg. Co., *Catalog* C.

VARISTORS AND THERMISTORS

By N. Y. Priessman

GENERAL. The term varistor, from the words variable and resistor, is applied to a group of circuit elements broadly classified as non-ohmic resistances. The application of the term varistor is restricted to devices in which the property of variable resistance is provided by solid semiconductor materials. The semiconductors that have proved useful as stable circuit elements are those in which the current carriers are electrons as distinguished from those in which ions are transferred through the solid.

These non-ohmic resistors may be divided into two broad classes, depending upon whether the resistance change is an electric field effect (varistors) or a temperature effect (thermistors). The field-effect varistors may further be divided into rectifier varistors and symmetrical varistors. Rectifiers, such as copper oxide, selenium, silicon, and germanium, exhibit quite different values of resistance depending on the polarity of the applied voltage. Symmetrical varistors such as Thyrite, Metrosil, Atmite, and silicon carbide show no rectifying properties. Thermistors (see article 10) change resistance markedly with changes of temperature but do not, independently of temperature change, possess a non-linear resistance characteristic.

8. COPPER-CUPROUS OXIDE VARISTOR

The copper-cuprous oxide varistor consists essentially of a piece of sheet copper about 0.050 in. thick in the form of a disk, washer, or plate which has been oxidized so as to form on its surface a layer of red cuprous oxide. A thin layer of conducting material is applied to the exposed surface of this oxide to provide a contact, known as the outer contact. The mother copper provides the other electrode. Figure 1 shows an enlarged and out-of-scale cross-section of such a varistor cell.

The manufacturing techniques, although differing in detail with various manufacturers, depending on the particular qualities in the product they are interested in, have in common the following: the desired form of cell is blanked from sheet copper (usually the grade known as Chile copper), chemically cleaned, oxidized in air atmosphere at a temperature in the neighborhood of 1000 deg cent for some 10 to 20 minutes to form a layer of cuprous oxide 0.003 to 0.004 in. thick, held for some minutes at a temperature of about 550 deg cent, and then quenched in water. At this stage the cuprous oxide (red) is covered with a thin layer of cupric oxide (black) which must be removed by chemical action. The means of providing an outer contact vary with the manufacturer. The following means are in common use: (*a*) painted-on contacts of "aqua-dag" (colloidal graphite), (*b*) electroplated contacts, and (*c*) contacts of gold or silver produced by the well-known techniques of evaporation in vacuum.

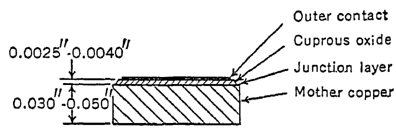


FIG. 1. Copper Oxide Varistor Cell

The electrical properties of the varistor may be altered significantly by suitable variations in the fabricating processes, as for example by the addition of metallic impurities to the copper (such as thallium); by changes in time, temperature, and atmosphere of the heat treatments; or by changes in rate of cooling and quench temperature, etc.

The current-voltage characteristics of a varistor cell commonly used in communication circuits is shown in Fig. 2 in the "Chile copper" curves. The "forward characteristic" exhibited by the cell with the mother copper negative and the outer contact positive is the flatter branch of the two curves. The "reverse characteristic" which obtains when the copper is positive and the outer contact negative is shown by the steeper branch. Considering in detail the forward characteristic curve the varistor behaves very much like an ohmic resistor at voltages below 0.05. As the voltage increases above this the current increases rapidly; that is, the resistance decreases. The seat of this potential dependent resistance as well as of the rectifying property is in the interface between the cuprous oxide and the copper. This interface is variously called blocking layer, junction layer, barrier layer, etc. As the voltage increases above 0.5 volt the current increases less rapidly and the resistance approaches a limiting constant series resistance, which is mainly the body resistance of the oxide layer itself. The series resistance of the outer contact is small compared with the body resistance of the cuprous oxide layer and is non-rectifying. The thallium-copper cell behaves similarly but with reduced currents in both directions.

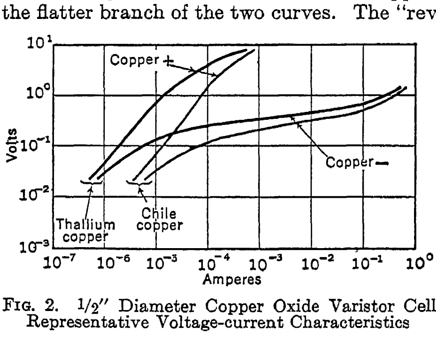


FIG. 2. 1/2″ Diameter Copper Oxide Varistor Cell Representative Voltage-current Characteristics

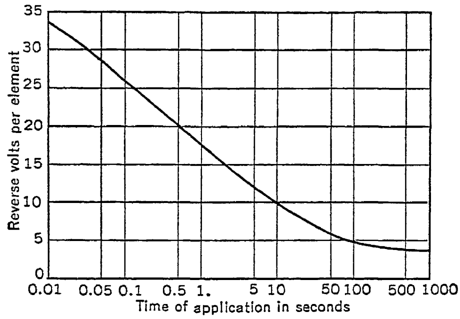


FIG. 3. Relation of Safe Reverse Voltage to Time of Application for Typical Copper Oxide Cells at Ambient Temperatures below 100° F and for Applications Spaced at Least 1 Minute Apart

The forward current at a voltage is proportional to the area of the cell or, more accurately, to the area of the outer contact. The reverse current is not simply related to the geometry of the cell. The characteristics of cells in series may be obtained by sliding the characteristic curves parallel to themselves along the voltage axis, and the characteristics of cells in parallel by sliding the curves along the current axis. In practice large areas may be obtained most economically by large-area plates.

Large deviations from the typical characteristic curves shown are inherent in the commercially manufactured product even when produced under the best controlled conditions of present practice. Moreover these characteristics change with time and use. While the forward resistance increases with time and temperature, the reverse resistance decreases with

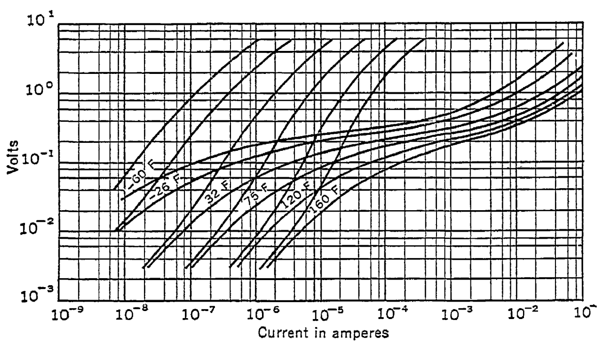


FIG. 4. Representative D-c Characteristic of $^3/_{16}$-in.-diam. Copper Oxide Varistor Cell (Chile Copper)

continued application of reverse voltage. In general this should not exceed 4 volts (see Fig. 3). These effects again vary from batch to batch.

Large negative temperature coefficients of resistance both in the forward and reverse direction are characteristic of these devices. No simple law such as holds for metallic conductors is applicable, and the variation of resistance with temperature may best be displayed graphically as in Fig. 4.

The effect of exposure to moisture is to reduce the reverse resistance of all types of cells. This is especially troublesome in the small-diameter, very-high-reverse-resistance cells. The forward resistance of cells with aqua-dag contacts is increased by exposure to moisture. It is customary in the use of varistors as circuit elements to provide substantial moisture-proofing in the form of organic coatings, potting in wax, etc.

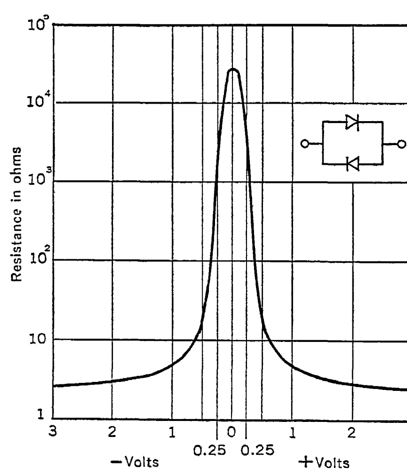


FIG. 5. Voltage Limiter, Two $^1/_2$-in.-diam. Copper Oxide Varistor Cells (Copper plus Thallium)

Structures. The $^3/_4$-in.-diameter and larger cells are usually made in the form of a washer and clamped together with wiring terminals on an insulated bolt. The $^1/_2$-in.-diameter and smaller cells are usually assembled with wiring terminals and pressure spring in cylindrical cavities in insulating blocks.

Ratings. When copper oxide varistors are used as current-supply rectifiers their ratings are based on four factors on which definite specifications should be made, with allowance for aging. These factors are (1) the d-c voltage and current output, (2) the a-c voltage input, (3) the ambient temperature, and (4) the means for cooling the varistor cells.

APPLICATIONS. Voltage Limiter. It may be seen from Fig. 2 that as the voltage in the forward direction increases from 0.1 to 1.0 volt the current increases more than 1000 fold. A varistor may be connected across the input terminals of a network to act as a bypass when applied voltages are substantially above the normal level. Figure 5 shows the resistance-voltage characteristic of such a varistor in which two 1/2-in.-diameter cells of thallium copper (see Fig. 2) are connected in parallel opposing so that the combination has a symmetrical resistance-voltage curve. The resistance is greater than 10,000 ohms for voltages up to 0.1 volt and drops to about 5.0 ohms at 1.0 volt. As the voltage increases above 1.0 volt the resistance decreases but little; it is being limited by the body resistance of the copper oxide layer.

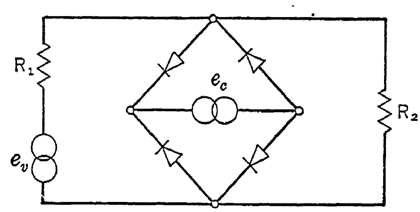


Fig. 6. Bridge Connected Modulator

Modulator. Copper oxide varistors are extensively used both as modulators and demodulators in carrier telephone systems. A bridge-connected modulator is shown in Fig. 6. The carrier voltage e_c is made large compared with the signal voltage e_v. When the carrier voltage is of such polarity as to bias the varistors in the forward direction they will all be low in resistance, offering substantially a short circuit to the signal current. When the carrier voltage reverses the varistors all have a high reverse resistance and the signal current appears in R_2. It is desirable to prevent unmodulated carrier current from appearing in R_2, and this is done by selecting the four varistors in the arms of the bridge to have voltage-current characteristics very closely alike so that the bridge is balanced throughout the cycle of carrier voltage. (A complete discussion of the varistor modulator is given by Caruthers, *Bell System Technical Journal*, Vol. 18, 315 [1939].)

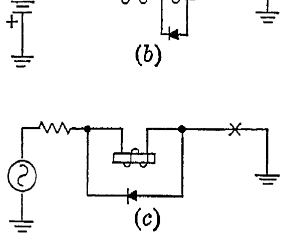


Fig. 7

Control of Telephone Relays. Various configurations of relay windings and varistors may be used to modify the response characteristics of relays. Several such arrangements are shown in Fig. 7: (*a*) delayed release with practically no effect on operate time; (*b*) delayed operate with practically no effect on release time. Steady operation of relays from an a-c source is obviously possible using the varistor as a rectifier. In the simple arrangement shown in Fig. 7(*c*), the reverse half-cycle is bypassed through the varistor which also affords a path for the slow decay of current established in the winding during the previous positive half-cycle.

Varistors are used in a circuit network known as a "compandor"—a contraction of the words "compressor" and "expandor." A compressor is a non-linear transmission network in which the range of signal power output is compressed relative to the range of power input. An expandor is a network which provides the inverse action. Detailed discussion of such circuits, also called "vario-lossers," may be found in Bennett and Doba, *Trans. A.I.E.E.*, Vol. 60, 17 (1941).

BIBLIOGRAPHY

A Bibliography on Metallic Rectifiers and Their Principal Applications, published by the American Institute of Electrical Engineering, New York, N. Y. A bibliography of more than 500 items on fundamental theory, various types of rectifiers and their applications.

Bibliography of Literature on Rectifiers and Semi-conductors, Royal Aircraft Establishment, Issue 4, August 1946.

H. C. Torrey and C. A. Whitman, *Crystal Rectifiers*, Rad. Lab. Series, McGraw-Hill Book Co.

9. SILICON CARBIDE VARISTORS

It has long been known that silicon carbide will, under suitable conditions of contact, exhibit a non-linear relationship between current and voltage. This may readily be demonstrated by measuring the voltage-current characteristic of a mass of small particles of silicon carbide compressed between metallic electrodes. As the voltage is increased from zero the current increases, at low voltage in direct proportion to the voltage and then much more rapidly. If the number of particles in the mass is large and the distance between electrodes large compared with the dimensions of the particles the non-linear resistance of the device is independent of polarity.

Experiments upon single particles with suitably made contacts indicate that the body resistance of the particle is small, ohmic, and independent of polarity.

The non-linear conduction exhibited by the mass of particles results from the voltage-dependent resistances at the point-to-point contacts between the granules of silicon carbide. The overall resistance characteristic may be thought of as due to large numbers of non-linear resistance contacts arranged at random in series and parallel. In a statistical sense the aggregate displays no dependence upon the direction of current flow. This varistor is an example of a "symmetrical non-linear resistor."

The simple device of containing a mass of silicon carbide particles under pressure between electrodes does not have the stability of characteristic under use conditions to afford wholly reliable circuit elements.

In 1930, McEachron (see *Journal A.I.E.E.*, Vol. 49, 410 [1930]) described a silicon carbide ceramic non-linear resistor to which the name Thyrite was given. The material consists of silicon carbide particles bonded in a ceramic matrix. Similar materials are known under various names such as Metrosil and Atmite.

The essential steps of manufacture are these: suitable silicon carbide particles, clay and water, sometimes with a minor constituent such as carbon, are mixed to form a plastic mass. The mass is partially dried and forced through screens to obtain a slightly damp granular powder. This material is compressed under high pressure into desired shapes, generally flat disks or rods. These pieces are further dried and heat treated in a reducing atmosphere at a temperature in the neighborhood of 1200 deg cent. The fired pieces are hard and strong and have mechanical properties quite similar to those of dry process porcelain. Electrodes on the opposite plane faces are provided by spraying or Schooping a layer of metal such as brass, copper, aluminum, or tin. The piece is then usually impregnated with a moisture-repellent organic substance to prevent pickup of water, which adversely affects their electrical stability.

The electrical properties of the product are profoundly affected by the parameters of process: materials, particle size, moisture content, forming pressure, and especially temperature, time, and atmosphere of the heat treatments. The products of different manufacturers differ somewhat in electrical properties, most importantly in the degree of non-linearity, and the characteristics of Fig. 8 are to be taken only as generally indicative. The current-voltage characteristic shown is closely represented by the equation

$$I = C_1E + C_2E^n$$

where I = current through the piece, E = voltage applied to the piece, C_1 and C_2 are constants depending on the material and geometry of the piece, and n is an exponent the value of which depends on various factors in the manufacturing process and generally lies between 3.5 and 5.0. Some manufacturers indicate values of n as high as 7.0 but only for pieces having resistances much above the range indicated in Fig. 8.

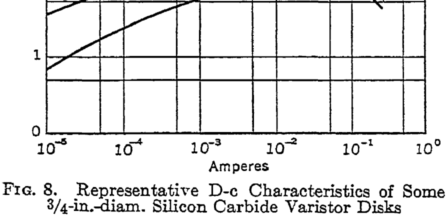


FIG. 8. Representative D-c Characteristics of Some 3/4-in.-diam. Silicon Carbide Varistor Disks

The variation of characteristic through control of manufacturing processes and geometry of the piece permits coverage of an enormous range of current and voltage. This range may be further extended by connection of pieces in series or parallel. It is to be noted from Fig. 8 that as the resistance of the piece decreases the value of n decreases also, and this being typical of all manufacturers' products may be considered an inherent characteristic of the presently made material. In consequence it is not possible with this device to obtain marked non-linearity at low voltage.

In common with semiconductors the silicon carbide varistor exhibits a negative temperature coefficient of resistance. The coefficient does not have a single value but varies both with the material and with voltage and temperature. The values of the coefficient at constant voltage cover a spread of from 0.3 per cent to 0.9 per cent per degree centigrade in the normally used range of temperature. The higher values of temperature coefficient are observed at the lower voltages.

At high frequencies consideration should be given to the presence of a capacitance effectively in parallel with the non-ohmic resistance. The exact value of this capacitance is determinable only by measurement, but the order of magnitude may be calculated by assuming the material to have a dielectric constant of 30 to 200.

Commonly used shapes are rods and disks. Small disks and rods may be furnished with leads soldered to the metallic electrodes on the faces of the piece. Disks are also made with holes in the center and clamped together with wiring terminals by means of a central bolt. Disks and rods of all sizes are used with spring clip mountings which furnish mechanical support and electrical connections.

When used under high humidity conditions, or at low currents, the organic impregnant, referred to in the description of the fabricating process, may not be sufficient protection against moisture and further precautions may be necessary.

Approximate values of mechanical and thermal properties of importance in circuit element design are as follows:

Bulk density....................	2.35 grams per cu cm
Compression strength............	15,000 to 23,000 lb per sq in
Specific heat....................	0.17 to 0.21 cal per gram per deg cent
Thermal conductivity............	0.0034 cal per cm per sec per deg cent

Requirements on the current-voltage characteristic for a particular application may be stated in a number of ways; the following are commonly used.

(*a*) The voltage E_1 at a current I_1 shall be greater than some value, and the voltage E_2 at a current I_2, where I_2 is greater than I_1, shall be less than some value. This statement of requirements contains implicitly a requirement as to the minimum value of n.

(*b*) The voltage at a given current I shall be equal to a value $E \pm X$ per cent, and the value of n shall lie within certain limits throughout a range of current.

It is to be noted that considerable differences in characteristic may exist between pieces meeting a set of such requirements. In commercial manufacture the range of voltage at a given current commonly runs ±20% about the average. Accuracy of meters used in checking requirements is important since errors in voltage readings are to be multiplied by n in determining their effect on current readings.

Self-heating resulting from power dissipation in the varistor lowers its resistance (negative temperature coefficient of resistance), but this effect is in general reversible; that is, no permanent effects on the characteristic are produced by moderate heating, say from 100 to 150 deg cent. The safe upper limit of heating is oftentimes determined by the moisture-resistant organic compound used as an impregnant. As shown in Fig. 8, 1.0 watt for a disk of $^3/_4$-in. diameter suspended in free air at 50 deg cent is a limit recommended by one manufacturer. Very heavy transient currents may alter permanently the characteristic, usually in the direction of decreasing the resistance.

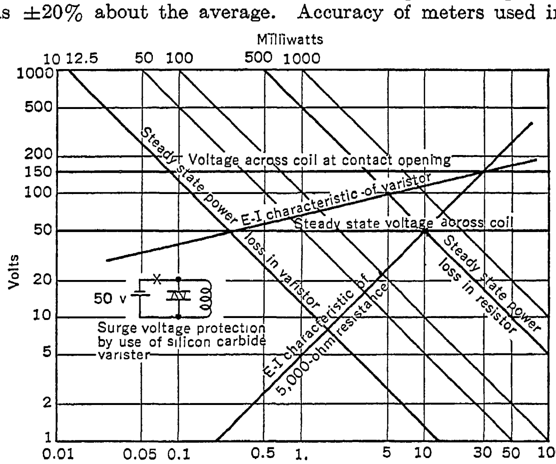


FIG. 9. Surge Voltage Protection by Use of Silicon Carbide Varistor

APPLICATIONS. (1) A silicon carbide varistor connected across the terminals of an electromagnetic winding acts to limit the surge voltage generated when the field is opened.

As shown in Fig. 9 the maximum value of voltage across the varistor may be determined from the point on the voltage-current characteristic corresponding to the steady-state value of current I_0 in the winding. As compared with an ordinary resistance shunt across the winding to secure the same voltage-limiting effect, the varistor dissipates much less power when the coil is steadily energized.

(2) In certain carrier telephone system filters exposed to high incoming voltage, the condenser of a high-Q combination of coil and condenser has been protected by a varistor in shunt.

(3) Some of the smaller telephone switchboards have line lamps connected directly in the subscriber's loop for signaling. These line lamps are exposed to electrical disturbances that may be impressed on the outside lines, and if the disturbances are severe enough the lamps may be burned out. Silicon carbide varistors have been used very effectively in parallel with the lamp to bypass large incoming surges. The high resistance of the varistor at the normal signaling level has no appreciable effect on the lamp illumination.

(4) Use is made of varistors to protect contacts controlling inductive circuits from the deleterious effect of sparks resulting from the opening of such circuits. Usually the varistor is connected across the winding rather than across the contact to avoid continuous current drain. Though such an arrangement is useful it is not a satisfactory general solution of the problem. The varistor increases the release time of the relay or switch magnet, though not to the extent that an ohmic resistance of equivalent spark quenching action would do, and it does not entirely eliminate high-frequency oscillations across the opening contact due to the associated wiring.

BIBLIOGRAPHY

McEachron, *J. A.I.E.E.*, Vol. 49, 410 (1930). (Thyrite.)
Fairweather, *J. A.I.E.E.*, Pt. I, Vol. 89, 499 (1942). (General. Contains bibliography of 59 items.)
Ashworth, Needham, and Sillars, *J. A.I.E.E.*, Pt. I, Vol. 93, 385 (1946). (General. Contains bibliography of 25 items.)
Brownlee, *Gen. Elec. Rev.*, Vol. 37, 175 (1934); Vol. 37, 218 (1934). (Calculation of circuits.)
Grisdale, *Bell Labs. Record*, Vol. 19, 153 (1941). (General.)
Vigren, *Telegrafstyrelsen*, No. 7, Vol. 9, 138 (1942). (Contact protection.)
Royal Aircraft Establishment Issue 4, August 1946, Bibliography of Literature on Rectifiers and Semiconductors.

10. THERMISTORS

Thermistors or thermally sensitive resistances are devices made of solid semiconductors the electrical resistance of which varies markedly with temperature. This phenomenon has long been known, Faraday having observed that the resistance of silver sulfide decreased rapidly as the temperature increased. Since that time it has been determined that a great number of materials classed electrically as semiconductors exhibit high negative temperature coefficients of resistance. Semiconductors have specific resistances at room temperature much greater than those of metallic conductors and much less than those of insulators. This very wide intermediate range of resistivities is not bounded precisely but may extend from 0.1 ohm cm to 10^9 ohm cm. Materials commercially employed in thermistor circuit elements have a much narrower range of resistivity, roughly from 10 ohm cms to 100,000 ohm cm.

The materials of thermistor construction include a wide variety of metallic oxides. In common use are the oxides of uranium and various mixtures of the oxides of magnesium, manganese, titanium, iron, nickel, cobalt, zinc, etc. The common method of fabricating is to heat the oxides in the form of compressed powders to a temperature at which they will sinter. At the sintering temperature the powders recrystallize to form a dense, hard, ceramic-like solid of homogeneous composition. The sintered-powder process permits the mixing of various oxides in suitable proportions to produce a wide range of electrical and thermal characteristics and permits as well the fabrication of a great variety of shapes and sizes of the completed piece.

Forms. Three forms of thermistors are common—disks, rods, and beads. A thin plate or flake form has also been described and is in limited use (see Becker et al., *Bell Sys. Tech. Jour.*, January 1947). Disks range in diameter from 0.125 to 2.0 in. and in thickness from 0.030 to 0.250 in. Rods are made in diameter from 0.030 to 0.250 in. and in length from 0.050 to 2.5 in. Bead diameters range from 0.006 to 0.060 in.

Properties. The relations between specific resistance and temperature of several thermistor materials are shown in Fig. 10 and for comparison the resistance-temperature relation of platinum. In Fig. 11 the log of the specific resistance is plotted against the reciprocal of the absolute temperature. It is seen that the curves are very nearly straight lines, and so to a close approximation

$$\log \rho = \text{const.} + \beta \cdot \frac{1}{T} \quad \text{or} \quad \rho = \text{const.}\ e^{\beta/T}$$

from which

$$\rho = \rho_0 e^{(\beta/T - \beta/T_0)}$$

where T = temperature in degrees Kelvin, $\rho = \rho_0$ when $T = T_0$, β is numerically proportional to the slope and is of the dimensions degrees Kelvin. From the definition of temperature coefficient $\alpha = (1/R)(dR/dT)$ it is seen that

$$\alpha = -\frac{\beta}{T^2}$$

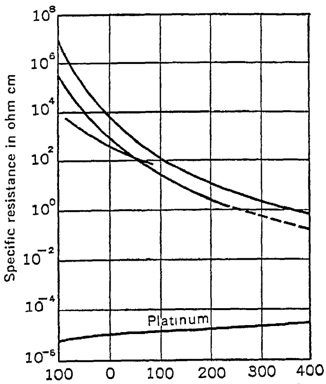


FIG. 10

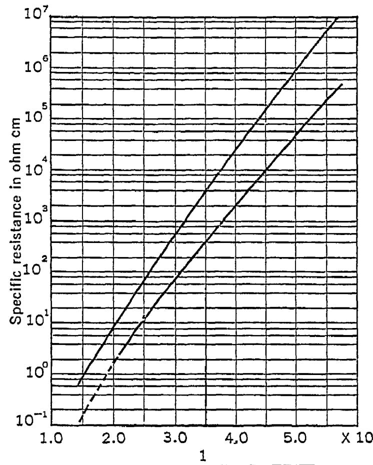


FIG. 11

THERMISTOR APPLICATIONS. Direct utilization of the resistance-temperature relation is a broad field of use including resistance thermometry, compensation for the positive temperature coefficient of other resistive circuit elements, temperature control, and the like. In all these applications the self-heating effect of any current in the thermistor is kept small so that the resistance is fully controlled by the ambient temperature.

The equation previously given for the relation between resistance and temperature may for purposes of general calculation be considered independent of temperature. Table 9 shows temperature-resistance characteristics of a typical thermistor thermometer. With an ordinary Wheatstone bridge and galvanometer and a suitably calibrated thermistor thermometer a precision of 0.001 cent deg is readily obtainable.

The use of thermistors in conjunction with relays, valves, etc., for temperature control is closely akin to their use in thermometry. The larger currents required for relay operation necessitate design consideration of the self-heating effects in the thermistor.

Thermistors are used to compensate for changes in resistance of electrical circuits caused by ambient temperature variations. Shunting the thermistor by a parallel resistance sometimes improves the accuracy of the compensation. Consideration should be given to like temperature exposure of the thermistor and the compensated circuit element, and also to the effects on both of power dissipation.

Table 9. Temperature-resistance Characteristic of a Typical Thermistor Thermometer

Temperature, deg cent	Resistance, ohms	Temperature Coefficients	
		B, deg cent	α, per cent per deg cent
−25	580,000	3,780	−6.1
0	145,000	3,850	−5.2
25	46,000	3,920	−4.4
50	16,400	3,980	−3.8
75	6,700	4,050	−3.3
100	3,200	4,120	−3.0
150	830	4,260	−2.4
200	305	4,410	−2.0

Dissipation constant in still air, approximately 4 milliwatts per degree centigrade; thermal time constant in still air, approximately 70 sec; dimensions of thermistor, diameter approximately 0.11 in., length approximately 0.54 in.

Small thermistors have been used extensively to measure power in very high-frequency test sets. Suitably mounted in a properly terminated waveguide a thermistor bead absorbs effectively the entering power, and the consequent heating of the bead produces a change in resistance which may readily be measured with high accuracy. Calibrating may be done with d-c or low-frequency power.

The self-heating effect of current through a thermistor, primarily the bead type, results in interesting and useful non-linear relationships between current and voltage. Figure 12 shows a "steady-state" characteristic of a particular form of bead thermistor. At small currents the power dissipated is too small to heat the thermistor appreciably, and the resistance remains constant. With increasing current the effects of self-heating become evident; the temperature of the thermistor rises and the resistance decreases. As the current continues to increase, the slope of the curve changes from positive to negative and in this latter region the thermistor exhibits a negative value of dv/dI, that is, a negative resistance. The numbers along the curve give the rise in temperature in degrees centigrade above the ambient. The change in resistance of the thermistor does not occur instantaneously with current change because of its thermal mass.

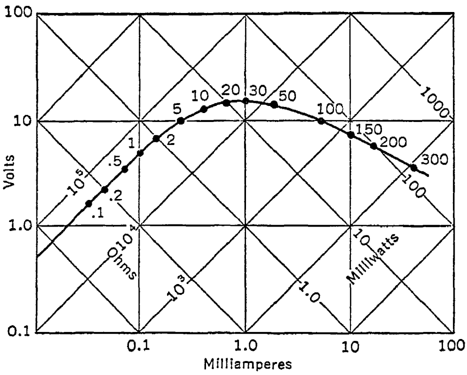


FIG. 12

The heating and consequent reduction of resistance by the continued passage of current is used to obtain delayed response circuits as well as non-response to short-duration surges by connection of the thermistor in series with a relay.

Figure 13 shows a combination of thermistor and resistances to obtain either a speech volume limiter or a volume compressor. The speed of response of the thermistor is adjusted to syllabic frequency or slower to eliminate the wave-form distortion and peak chopping common to quick-acting non-linear devices.

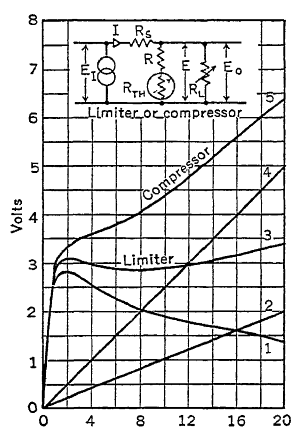


FIG. 13

Curve 1, thermistor characteristic
Curves 2 and 4, ohmic resistor characteristics
Curves 3 and 5, combined characteristics

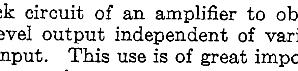


FIG. 14

Figure 14 shows a thermistor in the negative feedback circuit of an amplifier to obtain constant level output independent of variations of signal input. This use is of great importance in telephone carrier systems to correct for variations in overall line loss. The simple bead structure previously described is not adequate for this purpose since its resistance and hence the amount of feedback would be subject to change with changes in ambient temperature. Temperature compensation is economically obtained by associating a heater winding with the bead and regulating the input current to the heater to produce a constant temperature surrounding the bead.

BIBLIOGRAPHY

1. Pearson, G. L., Thermistors: Their Characteristics and Uses, *Bell. Labs. Rec.*, November 1940, p. 85.
2. Sillars, R. W., Materials and Devices of Falling Resistance-Temperature Characteristics, *J. Sci. Inst.*, Vol. 19, No. 6, 81 (June 1942).
3. Gray, T. S., and Van Dilla, M., Thermistors, Their Characteristics, Uses and Associated Circuits, Office of Publication Board Report PB-3407, 1945, Dept. of Commerce, Washington 25, D. C.

4. Gray, T. S., and Van Dilla, M., A Thermistor Electronic Thermoregulator, Office of Publication Board Report BP-3402 1945, Dept. of Commerce, Washington 25, D. C.
5. Gaffney, F. J., Microwave Measurements and Test Equipments, *Proc. I.R.E.*, Vol. 34, 775–793 (October 1946).
6. Becker, J. A., Green, C. B., and Pearson, G. L., Properties and Uses of Thermistors—Thermally Sensitive Resistors, *Electrical Engineering*, Vol. 65, November 1946, and *Bell Sys. Tech. J.*, Vol. 26, No. 1, 170–212 (January 1947).
7. Pearson, G. L., The Physics of Electronic Semi-conductors, *Trans. A.I.E.E.*, Vol. 66, 209 (1947).
8. Roloff, C. C., Thermo-variable Resistors, *Elec. Rev. Lond.*, Vol. 140, 315 (February 1947).
9. Verwey, Haayman, and Romeyn, Semi-conductors with Large Negative Temperature Coefficient of Resistance, *Phillips Tech. Rev.*, Vol. 9, No. 8, 239 (1947).
10. Montgomery, C. G., *Technique of Microwave Measurements*, M.I.T. Rad. Lab. Series. (Use of thermistors, Chapter 3.)
11. Deeter, E. L., Null Temperature Bridge, *Electronics*, Vol. 21, No. 5, 180 (May 1948).

INDUCTORS WITH AIR CORES

By L. M. Hershey

Of the many different types of air-core inductors in communications equipment today, the solenoid and the universal winding are most widely used. Torroidal windings, and other types such as the bank winding, spiral winding, and basket weave, are sometimes found.

The single-layer solenoid is used in untuned or tuned circuits, for resonant circuits, chokes, and in various other applications where a high Q, low distributed capacitance, mechanical strength, or ease of construction is of importance. More space is required to accommodate a single-layer solenoid than a universal winding of any type, for a given inductance; at frequencies below about 1 or 2 mc, space considerations frequently prevent the use of the single-layer solenoid.

The universal winding is popular in applications similar to those listed above for the single-layer solenoid but generally at frequencies below about 2 mc. It produces a coil having fairly high Q with low distributed capacitance, and mechanical strength. The distributed capacitance of a universal winding can be decreased by winding it on a narrower cam or by building it in a number of sections, each of which is a universal winding, with these sections connected in series. This multisection universal winding also can be adjusted to close inductance tolerance requirements by moving one of the end sections (sometimes called a "pi") nearer to, or away from, the adjacent section.

Many special types of windings are in use as tuned loops in the broadcast band (540 to 1600 kc). Among these types are the multilayer solenoid, basket weave, single-layer solenoid, and spiral winding.

Air-core inductors are used at frequencies up to about 200 mc, or up to a frequency where a transmission line becomes more convenient. The transition region between coils and transmission lines appears to be extremely broad.

The choice of the type of inductor is generally dictated by such practical considerations as available space or cost as well as by circuit considerations. Untuned primary coils and chokes of less than about 10 to 20 μh are usually solenoid windings. When greater inductance is required, the universal winding is used.

11. PROPERTIES OF AIR-CORE INDUCTORS

FIGURE OF MERIT. The figure of merit of an inductor is the ratio of its effective reactance to its effective resistance. This factor is called the Q of the inductor.

POWER FACTOR. The reciprocal of the Q of the inductor is equal to the power factor of the circuit within very close limits for values of Q above about 20. The power factor is more convenient than the Q to use in the calculation of certain circuit phenomena. For instance, the power factor of a circuit formed by an inductor shunted by a capacitor is the sum of the power factors of the two branches of the circuit, while the Q of the circuit is the reciprocal of the sum of the reciprocals of the Q's of each branch.

Like resistance and conductance, both the power factor and the Q are useful concepts, and the choice depends upon their application to the particular problem.

TIME CONSTANT. The time constant of an inductor in series with a resistor (the resistor may represent the internal resistance of the inductor) is L/R; it is the time in seconds required for the current, through an inductor of L henries in series with a resistor of R ohms, to reach 0.632 of its final value if a voltage is applied suddenly, or for the current through the series circuit to fall to 0.368 of its initial value if the inductor and resistor are short-circuited suddenly.

COIL LOSSES. The principal losses in an air-core inductor are those due to I^2R loss in the conductor and the dielectric losses in the coil form, wire insulation, impregnating material, etc. Eddy-current losses also occur in the conductor. At very high frequencies the losses due to radiated power may be appreciable. Additional losses occur outside the coil itself whenever any magnetic, dielectric, or conducting material is within the field of the coil.

The attainment of a maximum Q in a given space is one of the most common problems. At the lowest frequencies, the problem is to obtain the lowest d-c resistance for a given inductance. As the frequency becomes higher, skin effect (the tendency of an alternating current to flow along the outside surface of a conductor) becomes apparent, and is measurable even at power-line frequencies. Dielectric losses also begin to be noticeable at very low frequencies. Throughout the radio-frequency spectrum these two causes of power loss are of extreme importance. For example, a well-designed coil operating at 1 mc might have a d-c resistance of less than 5 ohms and an apparent resistance greater than 10 ohms.

Coil losses may be minimized, and high values of Q realized, by careful choice of the type of wire (to increase the surface area of the conductor through which the r-f currents flow), by obtaining the optimum spacing between conductors, and by choosing the proper shape of the winding for the available space.

It is sometimes desirable to design a coil with a certain value of Q so that, for instance, a required band width can be produced without the use of an external damping resistance. Then the designer may reverse his usual thought processes and use a "poor" shape factor for his coil, a conductor either larger or smaller than the optimum value for maximum Q, or use a value other than optimum for the spacing between conductors. Frequently, the diameter of the coil form can be reduced until a desired low value of Q is realized.

DISTRIBUTED CAPACITANCE. Each turn of an inductor is coupled magnetically to other turns of the same inductor. A certain small amount of capacitance between the turns of a winding is unavoidably produced by their proximity. The effect of all these small series capacitances across the whole inductor at its working frequency is called the distributed capacitance of the coil. The values of the distributed capacitances of various common types of air-core inductors range from a small fraction of a micro-microfarad up to 10 or more $\mu\mu$f. In general, a coil having a large ratio of length to diameter has a low distributed capacitance, and it is obvious that finer wire and greater spacing between turns will result in lower distributed capacitance.

Dielectric losses generally are decreased by a reduction in distributed capacitance. However, the changes necessary to reduce the dielectric losses and distributed capacitance (reducing the wire size, increasing the spacing between turns) finally begin to increase the copper losses in the conductor faster than the dielectric losses are decreased. An optimum design is a compromise between all these factors, and in a resonant circuit the distributed capacitance is frequently of lesser importance than the Q of the inductor; therefore, the capacitance is disregarded, while efforts are directed toward the attainment of minimum total losses. In r-f choke coil design, a minimum value of distributed capacitance is usually desired, while a Q higher than 5 or 10 produces a negligible effect. Therefore, choke coils are frequently wound with a minimum wire size, on long slender forms, and with relatively large spacing between turns.

12. ELECTRICAL DESIGN CONSIDERATIONS

TYPES OF CONDUCTORS. Copper wire is most commonly used in inductors. Copper tubing, or strips, are sometimes resorted to at frequencies above about 50 mc. Litz wire, composed of a number of strands of fine enameled wire (from about No. 38 to No. 44), produces lower r-f resistance than a single wire of the same area of cross-section. It is most effective in the lower-frequency part of the radio spectrum, below about 2 mc. Above this frequency the r-f currents appear to flow along the outside of the *group* of conductors and the effectiveness of litz wire is not so apparent.

Bare copper wire is rarely used for inductors. Tinned wire is found occasionally on space-wound solenoids; silver plating is also used occasionally on the heavier conductors normal for this type of coil. Silver plating offers the advantages of high conductivity on the surface of the conductor where higher-frequency currents flow, soldering to the conductor is made easy, and an attractive and fairly durable finish is produced.

TYPES OF WIRE INSULATION. Silk and cotton have been the most common materials for wire insulation on air-core inductors. Coil wire is made with a spiral wrapping, adding about 0.002 in. to the diameter of the wire for silk or about 0.004 in. for cotton. This is called a "serving" of silk or cotton. A double serving may be used, with the second serving spiraling in a direction opposite to that of the first serving. This is known as double-silk or double-cotton insulation; it is usually designated D.S. or D.C. For instance, a No. 38 bare wire with double silk insulation could be described as 38 D.S. Celanese and nylon are rapidly replacing silk and cotton as insulation for coil wire. Braided fabric insulation is also used to some extent.

An enamel coating is used, either alone on wire for solenoids, or with one or more servings of silk or one of the other serving materials over it. The enamel adds about 0.001 in. to the wire diameter. Plastic coatings can be added to conductors, and almost any required outside diameter can be produced in this manner.

Litz wire, as it is made in this country, consists of three or more strands of enameled wire with one or more servings of silk, cotton, celanese, or nylon, around the group of wires. The strands are either twisted in a regular fashion as they are wrapped or simply placed side by side parallel to each other. The twisted method produces a slightly higher Q under some conditions.

BEST COIL SHAPE. Formulas are available which express the best shape factor and winding pitch of a single-layer solenoid under idealized conditions. In a practical design problem these formulas serve as a valuable guide.

It has been shown that with a given length of wire, wound with a given pitch, the single-layer coil which has the maximum inductance value is so shaped that the ratio Diameter $\div$ Width of winding = 2.46, approximately.

Brooks has determined that there is a most efficient multilayer coil form to produce the maximum inductance with a given length of conductor. This most efficient inductor was produced as a compact multilayer cylindrical coil with a mean diameter 2.95 times the side of the square cross-section. Other proportions varying somewhat from the optimum affect the inductance only slightly. It has been determined that, when the ratio of mean diameter to side of square cross-section is 2.80, the resulting inductance is only 0.04 per cent less than the maximum value. It is generally convenient and within limits of accuracy to consider the optimum form as having the dimensions: diameter equal to 3 times width, and width equal height of winding.

Precision design can be effected only when the inductor is solely for low frequencies. The problem then of constructing an inductor of definite inductance for radio frequencies resolves mainly into the problem of minimizing the resistance and distributed capacitance. A coil is designed for a certain range of frequencies, and generally an attempt is made to construct a coil with a uniformly high value of Q in this range. In consideration of these and other requirements, a coil design will generally depart from the optimum proportions indicated above.

It will be found that a practical coil is, in general, somewhat elongated in the direction in which distributed capacitance will be minimized. For instance, a practical solenoid is usually longer in proportion than is indicated by the formula above; a diameter-to-length ratio of slightly over unity is common. The universal winding is frequently elongated in the radial direction. A diameter-to-length ratio of about 0.7 was found to produce the maximum Q over the broadcast band for a progressive universal winding on a 0.5-in.-diameter Bakelite form.

Dielectric and eddy-current losses, which usually are neglected in the computation of "best" coil shapes, appear to be primarily responsible for the noticeable discrepancies between calculated and measured data.

SOLENOID WINDINGS. Single-layer solenoids are sometimes wound with each turn touching the preceding turn; this results in fairly high values of distributed capacitance and eddy-current and dielectric losses. The length of the winding, and consequently its inductance, vary considerably from coil to coil in this type of winding because of variations in wire size (unless the wire is selected to closer than the usual limits for its gauge) and because of the failure of each turn to lie snugly against the next throughout its length. Therefore, when the inductance tolerance of a coil is closer than a few per cent, it is customary to "spin" a few turns—about 5 to 10 per cent of the total on the end of the coil. These turns are wound with the same spacing as the main portion of the winding, but the group of wires in the spun portion is spaced about $^1/_8$ in. away from the balance of the coil. These "spun" turns can be moved on the coil form nearer to, or away from, the balance of the winding.

It is possible to wind a solenoid to very accurate length and inductance limits (on a form of accurate diameter) on a winding machine or lathe adjusted to produce the desired pitch per turn. This method allows uninsulated wire to be wound in a solenoid, if care is taken to secure the turns properly on the form to prevent slippage after the coil is wound. Depending upon the design requirements, the pitch may be chosen to give the minimum spacing necessary to assure mechanical and electrical uniformity with maximum variations in wire and insulation dimensions, or the spacing may be chosen for optimum electrical performance.

Solenoids are sometimes wound in a screw-thread groove in the coil form, but unless a molded form is used it is difficult to maintain an accurate effective diameter of the groove and inductance variations result. Also, winding the conductor in a groove may reduce somewhat the Q of a very efficient inductor because of increased dielectric losses in the material in the immediate vicinity of the conductor. A slight increase in distributed capacitance will also result.

A simple manner in which to wind a single-layer solenoid is to choose insulation of the proper thickness to space the turns properly when the insulated wire is close wound.

THE UNIVERSAL WINDING. This type of inductor is wound in single or multiple sections ranging in widths from about $^1/_{16}$ to $^1/_2$ in. On a simple type of winding machine, the width of the coil is controlled by a cam which oscillates the wire guide back and forth in a linear fashion on the periphery of the form or on the next lower layer of the same winding. The cam is geared to the main shaft of the winding machine. The main shaft holds the coil form and rotates the coil as it is wound. The number of teeth in the gear on the main shaft (or driven by the main shaft through a 1/1 gear ratio) over the number of teeth in the gear on the camshaft is called the gear ratio. The wire is wound on the form at an angle to the side of the coil form (winding angle) determined by the width of the cam, the diameter of the coil, and the gear ratio. Simon has stated that a practical limit of this winding angle is about 12 deg maximum. Above this value, the wire may slip on the coil form and a poor winding will result. The winding angle becomes smaller as the coil builds up, being approximately inversely proportional to the diameter of the winding at any point; when this angle is reduced to about 6 deg the turns cross each other at an angle which is too small and tend to align themselves in the spaces between adjacent turns on the previous layer. The coil will not build up properly after this point is reached.

When a coil must be wound up to an outside diameter about equal to, or greater than, twice its inside diameter, it is frequently necessary to tolerate some slippage of the wire on the form at the start of the winding. The winding problem is, of course, greatly facilitated when the designer restricts the height of the winding to a point within practical limits.

The winding angle is proportional to the cam width and also to the gear ratio, since the gears drive the cam at a rate depending upon their relative number of teeth.

If g', the pattern gear ratio, is the fraction q'/s', where both q' and s' are small whole numbers, a simple and practical winding pattern should result. The number of cam cycles per winding cycles is q', and the number of planes cutting the periphery of the winding where interlaced crossing of turns occur is $(s' - 1)$.

In order to have the winding pattern repeat on consecutive winding cycles with the required spacing between adjacent turns, the pattern gear ratio may be corrected by the amount $\pm g'xw/2cq'$, where x is the desired number of wire diameters between turns (usually about 1.25), and w is the diameter of the wire. The plus sign produces a retrogressive winding; the minus sign produces a progressive winding. The gear ratio, g, may be computed $g = g'(1 \pm [0.63w/cq'])$.

THE PROGRESSIVE UNIVERSAL WINDING. The progressive universal winding is a special type of universal winding in which the wire guide is moved parallel to the axis of the coil form as the coil winds. The machine for winding this type of coil is usually equipped with 100:1 reduction gears driving the set of gears on the rear of the machine which produce the progression. The pitch of the winding produced by a 1:1 set of progression gears is usually very close to 0.01 in., but not exactly so on all machines.

It is necessary to correct the gear ratio (which would be used for an ordinary universal winding) slightly in order to allow for the change in spacing which results from the progression. This can be done by reducing g' by the amount $-0.5g_pp/c$, where g_p is the progression gear ratio and p is the pitch. This factor is usually rather small.

The steps required to determine the winding machine setup for this type of winding follow:

First, determine the number of turns per inch by dividing the desired number of turns by the required length. For a relatively wide cam and short winding, it is desirable to subtract one cam width from the required length for this calculation in order to allow for the tapering off of the winding at its ends. At this point it is best to examine the result and determine the number of layers of wire which will be built up radially on the coil. If the number is less than about 2.5 a solenoid would be preferable; if it is more than about 5 layers, the progressive winding may not build up satisfactorily, and a plain universal winding of one or more sections may be better.

The cam for a progressive winding is usually as narrow as possible, since the distributed capacitance and dielectric losses increase with a wider cam. A wider cam, however, decreases the steepness of the slope of the pile of wire upon which a turn must be wound and, consequently, enables a coil having more layers to be wound.

The value of g' for a progressive winding is usually less than for a universal, since it is unnecessary to make provisions for winding the coil up to a height comparable to the form diameter. In this case it is desirable to choose q' and s', again fairly small whole numbers, but values such as $^9/_5$ or $^7/_{11}$, where q' and s' are larger, produce a better pattern on the surface of the coil. Landon and Joyner's "composite" winding is obtainable when a fairly complicated pattern is obtained.

SHIELDING. The successful operation of many of the modern communication laboratory devices and radio receiving sets depends upon the effectiveness of the shielding between the various parts, and so, in high-gain amplifiers, leads, vacuum tubes, transformers, and tuning coils are all shielded. This process generally requires the placing of a metallic shielding container around the individual parts. Shielding is attempted against both electric and magnetic fields and is particularly necessary in circuits carrying high-frequency currents.

As has been pointed out before, whenever material is brought into the influence of the electric and magnetic fields of an inductor there is a transfer of energy to that material. Parts of the inductor are generally at a potential higher than that of the shield, which is usually at "ground" potential. With this condition there is added to the distributed capacitance of the winding more capacitance to "ground." This further complicates the calculation of the actual inductance of the coil at high frequencies.

Shields are generally constructed of non-magnetic or magnetic metals such as iron, zinc, copper, and aluminum. In the shield used for the shielding of a high-frequency magnetic field, the efficiency of the screen depends upon the eddy currents produced in the shield. The energy involved in the circulation of these eddy currents is drawn from the field of the shielded inductor. Magnetic shielding is therefore always accompanied by an increase in the effective resistance of the shielded inductor.

When the resistivity and thickness of the metal shield remain constant and the frequency of the alternating electromagnetic field varies, the shielding increases as the frequency increases, owing to the increased flow of eddy currents. For a given kind of metal at any specified frequency the shielding efficiency increases as the shield thickness is increased. Under these conditions, a certain thickness of shield introduces a maximum resistance into the shielded circuit. The thickness of the shield which gives maximum added resistance to the shielded circuit decreases as the frequency increases. If shielding is to be obtained by eddy currents they must be free to flow as they will, which requires that there be no imperfect joints or breaks in the shield. A short-circuited coil may be used as a shield since the current induced in it by the field will set up an opposing field and give a zero local resultant.

In an "open"-circuit electrostatic screen, no eddy currents will flow, and the shield may be used to prevent the alternating field from reaching an impure dielectric and thus producing a loss. In this application the shield reduces the effective resistance of the electrical circuit.

Since the eddy currents in a shield set up a magnetic field opposing the field of the inductor, it is evident that there will be a reduction of the net field surrounding the inductor winding. There is therefore a change in the effective coil inductance, which results in a decrease of the inductance value.

Many investigators have attempted to state quantitatively the magnitude of the screening effects on coil inductance and resistance. An idealized mathematical solution of the problem (given in the *Wireless Engineer*) replaces the ordinary cylindrical screen by a spherical one, and the cylindrical coil by a dipole of the same magnetic moment placed at the center of the sphere. The development is possible because it has been found that the exact shape of the screening can is not important, and this permits the mathematical use of a sphere with a diameter the geometric mean of the three coordinates of the can. The expression developed shows the reduction of the effective inductance of the coil to depend upon the frequency, material constants, the screen thickness, and a linear dimension representing the can size

$$L = L_0\left(1 - \frac{2}{3}\cdot\frac{V_c}{V_s}\frac{a}{K}\right) \tag{1}$$

in which V_c equals volume of the coil (winding section $\times$ length); V_s equals volume of screening can; K equals a constant less than 1 ($K = 0.7$ when coil length = coil diameter); L_0 equals actual inductance of the short solenoid; and a equals factor depending upon f, dimensions, permeability, and resistivity of the screen can (a = almost 1 for non-magnetic materials).

A general interpretation which may be made of this expression is that the effect of the screen on the coil inductance varies inversely as the diameter of the screen can to the cube power.

This conclusion has been stated by Hayman, and the results of a simple expression of the effect have been accurately checked by experiment. The approximate expression which has been given for short coils is

$$L_{\text{screened}} = L_{\text{unscreened}} \left(\frac{D^3 - d^3}{D^3}\right) \tag{2}$$

where D = screen diameter and d = coil diameter. As the coil length approaches half the can length there is more influence from the ends of the can. This condition requires the use of a correcting factor which is expressed as

$$\text{End correction} = \left[1 - \left(\frac{l \text{ coil}}{2l \text{ can}}\right)^2\right] \tag{3}$$

The calculations with these expressions for coils which do not exceed half the can dimensions have checked experimental measurements within less than 1 per cent.

The effect of the eddy-current flow in drawing energy from the inductor results in an increase of the effective resistance of the inductor. If skin effect in the shield is negligible and the eddy currents are uniformly distributed through the screen, the effective addition to the coil resistance has been stated as

$$R_s = \frac{3}{2\pi} T^2 A^2 \frac{\rho}{tr^4} \tag{4}$$

in which A = cross-section area of coil, square centimeters; T = number of coil turns; t = thickness of screen, millimeters; r = radius of screen can, centimeters; and ρ = resistivity of the can material, ohms per cubic centimeter. When skin effect in the shield forces a non-uniform distribution of eddy current the above expression is modified to take the form

$$R_s = 0.95 \times 10^{-4} \, T^2 A^2 \frac{\sqrt{f\rho}}{r^4} \tag{5}$$

An indication of the presence of skin effect in the shield is gained from the expression

$$t\sqrt{\frac{f}{\rho}}$$

When this factor is less than 5000, the skin effect is negligible.

13. MECHANICAL DESIGN CONSIDERATIONS

FORM MATERIALS. Some of the factors to be considered in choosing a form material are its mechanical strength, dielectric properties, coefficient of thermal expansion, machinability, moisture absorption, power factor at the operating frequency, and cost. Sometimes the operating-temperature requirements will not permit the choice of an otherwise desirable material, or the heat required to solder leads to lugs on the coil form may soften the coil form and loosen the solder lugs.

It has been found advisable to select an extremely stable material, whenever frequency drift requirements are severe, rather than an unstable material with compensation elsewhere in the circuit for inductance changes. Glass coil forms are sometimes possible where only a very small thermal expansion can be permitted. The various kinds of glass, with coefficients of linear expansion of about 3 to 9 parts per million per degree centigrade, are among the best materials now available. Steatite and mycalex also exhibit good temperature stability characteristics.

While phenolic materials have been popular because of their adaptability for use as form materials and their relatively low cost, their coefficient of linear expansion is only fair, being about 30 parts per million per degree centigrade.

The coefficient of thermal expansion of materials which are not homogeneous (such as commercial laminated Bakelite tubing) frequently has different values for the radial and axial dimensions. Average values of these two coefficients can generally be furnished by the manufacturer or may be measured directly.

The coil form is usually much more rugged than the winding, and therefore the temperature instability of a coil is usually due to its form. Copper wire expands at about the same rate as Bakelite under varying temperature conditions. However, copper wire can be wound under tension on one of the more stable form materials and a stable coil will result; the form material must be sufficiently stronger than the wire so that the form is not distorted by the pressure of the wire.

Absorption of water by the coil form may result in loss of mechanical strength, a change in the distributed capacitance of the inductor, and consequent detuning of the circuit, and may produce conditions favorable to electrolysis or fungus growth. The various methods of treating coils, such as impregnating, varnishing, or the application of fungicides, may delay these harmful results, but it appears that the most certain way to avoid them is to choose materials which will not absorb water, if such materials are available and fulfill the other design requirements.

A partial list of the more important properties of some of the more frequently used form materials is given herewith.

Typical Properties of Form Materials *

Form Material	Dielectric Constant	Power Factor, % at 1 mc	Maximum Temperature, deg cent	Coefficient † of Linear Thermal Expansion	Water Absorption, % in 24 hr	Machinability
Bakelite, molded..	6.0	4.0	120	30.0	0.2	Fair
paper base......	5.5	4.0	120	30.0	0.2	Good
Glass............	6.0	0.4	Over 500	8.5	0	Very poor
Pyrex..........	4.5	0.2	500	3.5	0	Very poor
Magnesium silicate	6.0	0.3	1000	7.0	0.02	Very poor
Mycalex..........	7.0	0.3	300	8.5	0.04	Very poor
Polystyrene.......	2.6	0.03	75	70.0	0.01	Good
Porcelain.........	6.5	0.7	1000	4.0	0.5	Very poor
Rubber, hard.....	3.0	1.0	65	75.0	0.01	Fair

* Values given are subject to considerable variation.

† In parts per million per degree centigrade.

IMPREGNATION OF INDUCTORS. Coil impregnation serves two principal purposes: first, the impregnating material tends to seal out moisture; second, the impregnating material improves the mechanical strength of the winding and holds it more firmly in place on the coil form.

Either before or during the impregnation of the coil, it is necessary to drive out of the winding and form any moisture that may be present. This is accomplished either by baking the coil at a temperature slightly higher than 100 deg cent before impregnating or by maintaining the impregnating material at such a temperature while the coil is being impregnated. The maximum limit of the temperature used during baking or impregnating is the temperature that will damage some part of the winding or its form or the impregnating material.

There are many types and mixtures of different types of waxes which are used to imimpregnate coils. Resins also are mixed with waxes to improve their characteristics. The melting and softening temperatures, and the hardness of the wax, are determined by the kinds and amounts of the various kinds of waxes in the mixture. Some mixtures become extremely brittle at fairly low temperatures, which may be a serious disadvantage.

Varnish of good electrical quality, or polystyrene dissolved in a solvent, are sometimes used to impregnate coils. Vacuum impregnation (dipping the coil in the impregnating material while it is in a partial vacuum) is usually resorted to when varnish or some similar material is used to impregnate a coil.

The conditions under which the coil will be used determine the designer's choice of the method and material for impregnating the coil. Very often the economy of the method is a determining factor.

Coils which are designed for good conditions of temperature and humidity, such as in equipment to be used indoors in all but the most humid parts of the United States, are impregnated usually in a simple and inexpensive way. For instance, many coils are treated by baking in an oven until most of the moisture is driven out, followed by immersion in a wax, resin, or mixture of both, at a temperature slightly higher than the boiling point of water until all agitation of the liquid ceases. The coil is then cooled and finally "flash-dipped" to produce an even coating of the same or a different impregnant on the outer surface. The first (baking) operation is frequently omitted for the sake of economy.

Where more severe conditions of temperature and humidity are encountered, some better methods of impregnation and better materials are required. Two such impregnation procedures are described below. These methods were developed and tested by the Hazeltine Corporation for use on equipment for the Army and Navy during 1945. The two processes are:

Process 1. Q-Max A-27 diluted 1:1 with Toluene, applied by dipping and baking; recommended for Silicone-varnished Fibreglas-served wire.

(*a*) Bake the coil (on its form) for 1 hr at 110–120 deg cent.
(*b*) While still hot, immerse the coil until bubbling ceases in a solution consisting of:

1 part Q-Max Lacquer A-27
1 part Toluene ("technical grade")

(*c*) Drain and air dry for 1 hr.
(*d*) Bake for 4 hr at 140 deg cent.
(*e*) Apply two more coats per (*b*), (*c*), and (*d*) above, redipping immediately after (*d*).

Both processes described here require the use of combustible materials. Adequate ventilation must be provided and safety rules regarding fire hazards observed.

Process 2. Styrene Monomer N-100 and Q-Max A-27 diluted 1:1 with Toluene, applied by dipping and baking. Recommended for use on Fibreglas- or silk-served wire.

(*a*) Bake the coil (on its form) for 2 hr at 110 deg cent.
(*b*) While still hot, immerse the coil in styrene monomer for 20 min.
(*c*) Air dry the coil until dripping ceases.
(*d*) Bake for 24 hr at 125 deg cent.
(*e*) Redip in styrene monomer for 20 min.
(*f*) Repeat (*c*) and (*d*).
(*g*) Apply a thin coat of the Q-Max and Toluene solution.

Coils impregnated according to the two processes given above should withstand temperatures ranging from -65 to $+85$ deg cent with the relative humidity as high as 95 per cent at the highest temperature.

THE SPECIFICATION OF INDUCTORS. The materials and the method of construction of inductors can be shown readily by means of drawings and written specifications. However, it is sometimes difficult to specify the performance requirements of an inductor in an exact fashion. This is due to a lack of standardized measuring equipment in the industry capable of separating the effects of the distributed capacitance, inductance, and Q of the inductor in a practical and accurately measurable fashion, especially on the higher-frequency coils. It is possible, however, to compare one inductor to another with a higher degree of precision. Therefore, it has become a fairly widely accepted practice for the designer to adjust accurately one complete set of coils, which become the master standards. From these master standards, as many secondary standards as are required can be made and distributed to the coil adjusting and testing points. All performance specifications requiring very close tolerances are referred to these standards. If the standards are carefully prepared, stored, and handled, the variations in their performance over a period of time due to aging, etc., are minimized. It is desirable to obtain measurements, on the most stable equipment available, of all possible performance data of the standards so that the master standards themselves can be rechecked. If both the coil manufacturer and designer can measure the inductor, with similar equipment and by similar methods, the data and the inductors measured can be exchanged and the test equipment calibrated alike in both places. Even if this can be done, in view of the difficulty involved in obtaining standardized test equipment and conditions it is generally desirable to use the master standard as the basis of the specification of inductor performance.

14. INDUCTOR DESIGN FORMULAS

The formulas that follow have been found to be generally useful in the design of inductors. The inductance formulas are less accurate than those given in the *Bureau of Standards Circular* C74. They do, however, produce a result with approximately the same accuracy with which the usual range of radio-frequency inductors can be wound, using materials which are not selected to closer than normal limits. Their principal advantage is their simplicity.

The dimensions and symbols used in these formulas follow:

L = inductance of each section of a winding in microhenries.
d = form diameter in inches.
b = length of winding in inches.
c = throw of cam used to wind a universal section in inches.
w = outside diameter of wire, including covering insulation, in inches.
t = number of turns.
C_0 = distributed capacitance in micro-microfarads.
p' = pitch of winding in inches.
h = height of winding above coil form in inches.
a = mean diameter of winding in inches.

L_t = total inductance of a multisection universal winding in microhenries.
g_p = progression gear ratio.
y = number of sections in winding.
g = gear ratio.
g' = pattern gear ratio.
q' = small whole number, numerator of fraction g'.
s' = small whole number, denominator of fraction g'.
z = number of strands in litz wire.
L_a = apparent inductance in microhenries of coil with distributed capacitance.
ω = 2π times frequency in megacycles per second.
p = pitch produced by winding machine when 1/1 progression gear ratio is used.

SOLENOID AND PROGRESSIVE UNIVERSAL WINDING. Inductance.

$$L = \frac{d^2t^2}{18d + 40b} \tag{6}$$

Accurate to within about 1 per cent for solenoids with $b > 0.4d$.

Turns.

$$t = \frac{\sqrt{L(18d + 40b)}}{d} \tag{7}$$

Distributed Capacitance.

$$C_0 = \frac{2.2d}{\cosh^{-1} p'/w} \tag{8}$$

for single layer solenoids only, where w is the diameter of the bare wire. The dielectric constant of the coil form, wire insulation, and impregnating material will increase the distributed capacitance above the calculated value.

For progressive universal windings, the distributed capacitance is the minimum for the smallest cam throw; it is approximately proportional to the coil diameter.

EACH SECTION OF A UNIVERSAL WINDING. Inductance.

$$L = \frac{0.2d^2t^2}{3d + 9c + 10h} \tag{9}$$

Accuracy to within about 1 per cent when the three terms in the denominator are about equal.

Turns.

$$t = \log^{-1} (1.08 - 0.47d + 0.16c + 0.5 \log L) \tag{10}$$

to within 5 per cent approximately, for windings having more than about 100 turns. This is an empirically derived formula. If more accurate results are required, the result may be substituted in eq. (6) above, and a suitable correction obtained.

Height of Winding.

$$h = \frac{1.25w^2t}{c} \tag{11}$$

for wire which does not flatten appreciably during winding, and with 1.25-wire-diameter spacing. For litz wire, flattening of the insulation may reduce the calculated height by about 5 to 10 per cent.

Distributed Capacitance. The distributed capacitance of a universal winding is extremely variable; it depends upon many factors, including the pattern gear ratio (g'), the spacing between conductors, etc. It is approximately proportional to the throw of the cam and to the form diameter. It is fairly independent of the number of turns on the winding.

MULTISECTION UNIVERSAL WINDING. Inductance.

$$L_2 = L(2y - 1) \tag{12}$$

approximately, for each section with uniform spacing between sections about equal to the width of the section.

Distributed Capacitance. Slightly greater than the resultant of the distributed capacitances of the individual sections in series.

SPIRAL WINDING. Inductance.

$$L = \frac{(d + h)^2t^2}{16d + 44h} \tag{13}$$

where h is the difference between outside and inside diameters.

MEASUREMENT FORMULAS. Apparent Inductance.

$$L_a = \frac{L}{1 - \omega^2 L C_0} \tag{14}$$

Distributed Capacitance.

$$C_0 = \frac{C_1 - 4C_2}{3} \tag{15}$$

where C_1 and C_2 are the capacitances required to resonate the inductor at frequencies f and $2f$, respectively. When C_1 and C_2 are large, the value of C_0 obtained will be somewhat larger than that computed from the actual self-resonance of the coil.

FIGURE OF MERIT. The Q of an inductor is usually measured directly on equipment designed for that purpose. If such equipment is not available, a voltage can be induced in the inductor in a resonant circuit, with suitable precautions taken to avoid additional loading, and the resonant band width measured at the half-power point. The ratio of the test frequency to the distance between the half-power points is equal to the Q of the inductor.

UNIVERSAL WINDING GEAR RATIO CALCULATIONS

Minimum cam width = Approximately $3w$

Cam Cycles per Winding Cycle.

$$g' = \frac{q'}{s'} = \frac{d}{3c} \tag{16}$$

for forms of average smoothness; this was expressed first by Simon in terms of cross-overs (one-half cam cycle) per winding cycle:

$$g' = \frac{q'}{s'} = \frac{d}{4c} \tag{17}$$

for very smooth forms. Choose a small whole number for q' and s', with

$$s' < \frac{c}{3w} \tag{18}$$

Gear Ratio.

$$g = g'\left(1 \pm \frac{0.63w}{cq'}\right) \tag{19}$$

Set the calculated value of g on the C or D scale of a 10-in. slide rule, and locate the coincidence of two whole-number lines within the range of available gears on the C and D scales. These two coincident line numbers can be used as the number of teeth in the gears.

PROGRESSIVE UNIVERSAL WINDING GEAR RATIO CALCULATION. To determine whether progressive universal should be used, if

$$\frac{w}{g_p p} < 1 \tag{20}$$

a solenoid should be used, according to Landon and Joyner. To determine minimum cam throw:

$$c > \frac{4w^2}{g_p p} \tag{21}$$

which is Landon and Joyner's formula in the terms used here. To determine cam cycles per winding cycle:

$$g' = \frac{d}{4c} = \frac{q'}{s'} \tag{22}$$

Choose fairly small whole numbers for q' and s'; the more complicated patterns result when s' is made fairly large, and a proper choice of q' and s' can be made so that the wire is adjacent to a preceding turn in the pattern on both forward and backward strokes of the cam, producing Landon and Joyner's "composite" winding. To determine progression gear ratio:

$$g_p = \frac{b}{tp} \tag{23}$$

To determine gear ratio:

$$g = \left(g' - \frac{0.5 g_p p}{c}\right)\left(1 \pm \frac{0.63w}{cq'}\right) \tag{24}$$

This is essentially the original formula published by Landon and Joyner, in the form suggested by Simon. Use the minus sign after g' when the winding forms a right-hand screw thread. If a left-hand thread is formed, a plus sign should be used. The value of g may be computed in the same manner as for a universal winding.

BIBLIOGRAPHY

Handbook of Radio Standards, National Electrical Mfr's. Assoc., New York (1928).

Radio Instruments and Measurements, *Bureau of Standards Circular* 74 (1924).

Morecroft, J. H., *Principles of Radio Communication*, John Wiley (1933).

Grover, F. W., Additions to the Formulas for the Calculation of Mutual and Self Inductance, *Bureau of Standards Scientific Paper* 320 (1918).

Grover, F. W., Formulas and Tables for the Calculation and Design of Single Layer Coil, *Proc. I.R.E.*, Vol. 11, 193 (April 1924).

Grover, F. W., Comparison of the Formulas for the Calculation of the Inductance of Coils and Spirals Wound with Wire of Large Cross Section, *Bureau of Standards Journal of Research*, Vol. 3, No. 1, 163, *Research Paper* 90 (July 1929).

Snow, C., Formulas for the Inductance of a Helix Made with Wire of Any Section, *Bureau of Standards Scientific Paper* 537 (1926).

Grover, F. W., Methods for the Derivation and Expansion of Formulas for the Mutual Inductance of Coaxial Circles and for the Inductance of Single Layer Solenoids (includes extensive bibliography), *Bureau of Standards Journal of Research*, Vol. 1, No. 4, 487, Reprint 16 (October 1928).

Snow, C., The Effect of Small Variation in Pitch upon the Inductance of a Standard Solenoid, *Bureau of Standards Journal of Research*, Vol. 6, No. 5, 777, *Research Paper* 304 (May 1931).

Brooks, H. B., Design of Standards of Inductance and the Proposed Use of Model Reactors in the Design of Air Core and Iron Core Reactors, *Bureau of Standards Journal of Research*, Vol. 7, 289, *Research Paper* 342 (August 1931).

Rosa, E. B., and Cohen, Louis, Formulae and Tables for the Calculation of Mutual and Self Inductance, *Bulletin of the Bureau of Standards*, Vol. 5, No. 1, 1, *Tech. Reprint* 93 (August 1908).

Rosa, E. B., and Grover, F. W., Formulae and Tables for the Calculation of Mutual and Self Inductance, *Bulletin of the Bureau of Standards*, Vol. 8, No. 1, 1, *Reprint* 169 (January 1912).

Snow, C., Simplified Precision Formula for Inductance of a Helix with Lead-in Correction, *Bureau of Standards Journal of Research*, Vol. 9, 419 (September 1932).

Wheeler, H. A., Simple Inductance Formulas for Radio Coils, *Proc. I.R.E.*, Vol. 16, No. 10, 1398 (October 1928).

Brainerd, J. G., Inductance at High Frequencies and Its Relation to the Circuit Equations, *Proc. I.R.E.*, Vol. 22, No. 3, 395 (March 1934).

Palermo, A. J., Distributed Capacity of Single Layer Coils, *Proc. I.R.E.*, Vol. 22, No. 7, 897 (July 1934).

Palermo, A. J., Effect of Displacement Current on Radio Frequency Resistance of the Single Layer Coil, *Proc. I.R.E.*, Vol. 20, No. 11, 1807 (November 1932).

A Critical Review of Literature on Amplifiers for Radio Reception, *Radio Research Special Report* 9, Department of Scientific and Industrial Research, published under the authority of his Majesty's Stationery Office, London (1930).

Breit, G., Some Effects of Distributed Capacity between Inductance Coils and the Ground, *Bureau of Standards Scientific Paper* 427, *Scientific Papers*, Vol. 17, 521 (December 1921).

Hickman, C. N., Alternating Current Resistance and Inductance of Single Layer Coils, *Bureau of Standards Scientific Paper* 472, *Scientific Papers*, Vol. 19, 73 (May 1923).

Breit, G., The High Frequency Resistance of Inductance Coils, *Bureau of Standards Scientific Paper* 430, *Scientific Papers*, Vol. 18, No. 17, 569 (1922).

Butterworth, S., Effective Resistance of Inductance Coils at Radio Frequency, *Experimental Wireless*, Vol. 3, 203, 267, 417, 483 (1926).

Austin, B. B., Effective Resistance of Inductance Coils at Radio Frequency. An abstract of a paper by. S. Butterworth (see above), *Wireless Engineer*, Vol. 11, No. 124, 12 (January 1934).

Palermo, A. J., and Grover, F. W., Study of High Frequency Coil Resistance, *Proc. I.R.E.*, Vol. 18, No. 12, 2041 (December 1930); Vol. 19, No. 7, 1278 (July 1931).

Hund, August, *High Frequency Measurements*, International Series in Physics, McGraw-Hill (1933).

Hund, A., and DeGroot, H. B., Radio-frequency Resistance and Inductance of Coils Used in Broadcast Reception, *Bureau of Standards Technical Paper* 298, Vol. 19, 651 (1925).

Hall, E. L., Resistance of Conductors of Various Types and Sizes as Windings of Single Layer Coils at 150 to 6000 Kilocycles, *Bureau of Standards Technical Paper* 330, Vol. 21, 109 (1926).

Barden, W. S., and Grimes, D., Coil Design for Short Wave Receivers, *Electronics*, Vol. 7, No. 6, 174 (June 1934).

Brown, W. W., and Love, J. E., Design and Efficiencies of Large Air Core Inductances, *Proc. I.R.E.*, Vol. 13, No. 6, 755 (December 1925).

Effect of Screening Cans on Effective Inductance and Resistance of Coils, *Wireless Engineer*, Vol. 11, No. 126, 115 (March 1934).

Hayman, W. G., Inductance of Solenoids in Cylindrical Screen Boxes, *Wireless Engineer*, Vol. 11, No. 127, 189 (April 1934).

Pollack, D., and Hartley, M., *Short Wave Coil Characteristics*, Research Labs., Columbia University (1935).

Simon, A. W., Winding the Universal Coil, *Electronics*, Vol. 9, 22 (October 1936).

Joyner, A. A., and Landon, V. D., Theory and Design of Progressive Universal Coils, *Communications*, Vol. 18, 5 (September 1938).

Hershey, L. M., The Design of the Universal Winding, *Proc. I.R.E.*, Vol. 29, 442 (August 1941).

Simon, A. W., On the Winding of the Universal Coil, *Proc. I.R.E.*, Vol. 33, 35 (January 1945).

Simon, A. W., On the Theory of the Progressive Universal Winding, *Proc. I.R.E.*, Vol. 33, 868 (December 1945).

Terman, F. E., Some Possibilities for Low-loss Coils, *Proc. I.R.E.*, Vol. 23, 1069 (September 1935).

Reber, G., Optimum Design of Toroidal Inductances, *Proc. I.R.E.*, Vol. 23, 1056 (September 1935).

Pollack, D., The Design of Inductances for Frequencies between 4 and 25 megacycles, *RCA Review*, Vol. II, 184 (October 1937), and *Electrical Engineering*, September, 1937.

FERROUS-CORED INDUCTORS

By A. J. Rohner

The utility of ferromagnetic cores in coils lies in the fact that such cores have a higher magnetic permeability than air. This permeability may be anywhere between 2 and 100,000. The use of a ferrous core may have the following beneficial effects:

(*a*) Increase of inductance. With a complete magnetic path of ferromagnetic material, this increase may be several thousand times the air-core value of inductance.

(*b*) Increase of Q. This results from the increased inductance, if the increase of loss due to core loss is not greater than the increase of inductance.

(*c*) Magnetic shielding. The magnetic field of the coil is constrained to follow to a large extent the path of the high-permeability core.

(*d*) Adjustability of inductance. Movement of the core in or out of the coil, or variation of an air gap in the core, gives mechanical means of adjusting the inductance.

The limitations of ferromagnetic cores in coils are due to certain undesirable qualities of ferromagnetic materials. Most important of these are:

(*a*) Magnetic saturation, which occurs at 7000 to 15,000 lines per square centimeter. These values depend upon the kind of core material and are for a magnetizing force of 10 oersteds.

(*b*) Non-constant permeability. Permeability varies with the direct current passing through the coil, the alternating voltage impressed across the coil terminals, and other factors.

(*c*) Core loss, which is a wattage loss additional to the copper loss of the coil; it determines the frequency range for which each type of core material may be used.

Core loss consists of two distinct parts, hysteresis loss and eddy-current loss. See Section 2, "Magnetic Materials," Spooner, "Properties and Testing of Magnetic Materials," or "Magnetic Circuits and Transformers" by staff of M.I.T. Hysteresis loss is a magnetic effect due to the magnetizing and demagnetizing of the core and is proportional to the frequency. Hysteresis loss can be reduced by using core material that is easily magnetized and demagnetized, i.e., a "magnetically soft" material. Eddy-current loss is an electrical effect due to induced currents within the core material and is proportional to the square of the frequency. Eddy-current loss can be reduced by using core material of high electrical resistivity, and by laminating or powdering the core.

At the higher frequencies, eddy-current loss becomes predominant and a finer subdivision of the core material is necessary.

Frequency Ranges. At frequencies below about 4 kc, the core usually consists of thin sheets, either in the form of flat plates or "laminations," as thin as 0.003 in., or in the form of ribbon, which can be made as thin as 0.001 in. Above 20 kc, the core usually consists of powdered material, the grains of which may average as little as 0.00012 in. in diameter. Between 4 and 20 kc, either sheet material or powdered material may be suitable. However, thin ribbon cores are useful up into the low- and medium-radiofrequency bands, while powdered cores, of larger grain size, are useful down to 1 kc or lower.

15. LOW-FREQUENCY, SHEET-CORE INDUCTORS

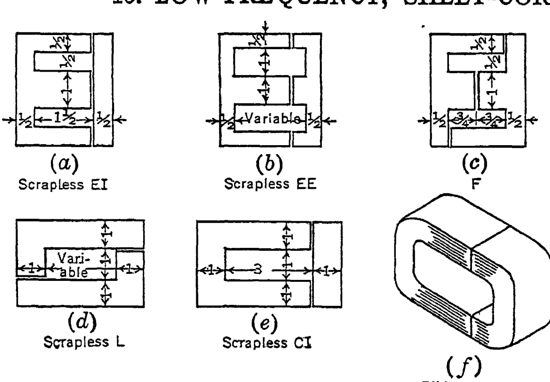


Fig. 1. Laminations and Ribbon Core

CORE CONSTRUCTION. Typical lamination shapes, and a ribbon core, are shown in Fig. 1. In general, all sheet-material cores form a complete magnetic path around the coil, except for small air gaps which may be purposely introduced, and the cross-section of this magnetic path is essentially uniform throughout the length of the magnetic circuit.

Assembly of laminated cores is illustrated in Fig. 2. A butt joint is actually a small air gap, because of oxide on the ends of the laminations, non-squareness of the ends, and imperfect meeting of the joint. This butt-joint gap may be from 0.0005 to 0.002 in. long. A value of 0.0015 in. per gap may be assumed for average design purposes. When there is an air gap, the magnetic flux usually must cross this separation *twice*. Thus, for example, if the air-gap spacer is 0.010 in. thick, the total length of air gap in the core, including the butt-joint gaps, is 0.023 in.

Usually, ribbon cores are cut in two after winding, forming two C-shaped pieces, and the ends of each C-piece are ground flat. When the two pieces are placed together, after adding the coil, the ends meet in tight butt joints, each joint being about 0.0005 in. long, so that the core as a whole approaches a continuous ferromagnetic path.

Stacking Factor. Because of oxide or other insulation on the sheet material, non-flatness of sheets, and stamping burrs on laminations, the magnetically useful cross-section of a sheet core is never 100 per cent of the measured cross-section. The ratio of the two is called the "stacking factor." Values that may be used as a guide are given in Table 1.

Table 1. Stacking Factors

Laminated Cores		Ribbon Cores	
19-mil.	0.94	14-mil.	0.65
14-mil	0.92	5-mil.	0.91
6-mil.	0.83	3-mil.	0.86
3-mil.	0.71	2-mil.	0.80

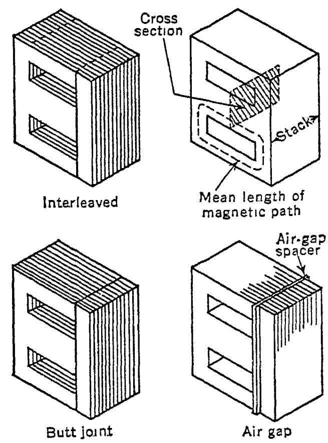


FIG. 2. Assembly of Laminated Cores

Materials most commonly used in sheet-material cores are:

(*a*) Silicon steel, especially the better grades, having silicon content from 2.5 to 4.75 per cent.

(*b*) Grain-oriented silicon steel. Hipersil and Silectron are trade names for this material.

(*c*) Nickel-iron alloys, of approximately 50–50 composition, variously known as Nicaloi, Hipernik, 4750, or 49-Alloy.

(*d*) Permalloy, an alloy of about 80 per cent nickel with iron. Hymu is a similar material.

(*e*) Mumetal, similar to permalloy, but with 5 per cent copper added.

For more detailed description of sheet-core materials, see Section 2, *Magnetic Materials*, Spooner, Chapter IV; Elmen, "Magnetic Alloys of Iron, Nickel, and Cobalt," *J. Franklin Inst.*, May 1929; Alleghany Ludlum *Bulletins* EM-11 and EM-12, and their *Magnetic Core Materials Practice;* Follansbee, *Electrical Sheet Handbook*, Magnetic Metals *High Permeability Alloys;* and Westinghouse, *Metals and Alloys*.

Thickness of Sheets. For inductors operating at 25 to 120 cycles, 25-mil-thick (U.S.S. gage 24) and 19-mil-thick (26-gage) laminations are useful. Sheet of 14-mil thickness (29 gage) is widely used, both for laminations and ribbon, for applications from 60 cycles to the middle audiofrequencies. Below 14 mils, sheet material can be obtained in almost any thickness down to 1 mil. However, stamped laminations, and preformed ribbon cores, of these thinner sheets became available largely as a result of war needs and are not yet standardized. Laminations can be purchased, of 7-mil silicon steel, 6-mil 49-Alloy and Hymu, 4-mil 4750, and 3-mil 4750, in a large variety of sizes. Hipersil ribbon cores are obtainable in 14-, 5-, 3-, 2-, and 1-mil sheet thickness.

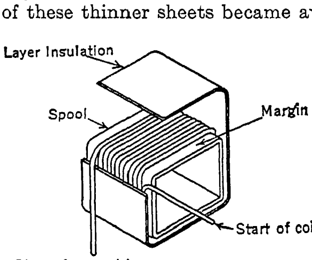


FIG. 3. Layer-wound Coil

COIL CONSTRUCTION. Low-frequency inductors usually have multilayer coils, with insulation between layers. See Fig. 3. The coil is wound upon a rectangular spool of spirally wrapped paper or fiber, 30 or 40 mils in total thickness, slightly larger in inside dimensions than the core over which it is to be placed, and slightly shorter than the core window. These clearances may be $^1/_{32}$ or $^1/_{16}$ in. Wire is usually solid copper, with enamel insulation. The length of the winding, or "wire traverse," is less than the length of the spool, to allow a "margin" of $^3/_{32}$ to $^3/_{16}$ in. at either end. Over each layer of wire is placed one turn of insulation, the same width as the length of the spool, which forms a smooth support for the next layer. Kraft paper, having a thickness about $^1/_5$ of the wire diameter, is a very satisfactory material for layer insulation. Glassine paper, of about the same thickness, is often used with wire of 24-gage (A.W.G.) or smaller. Layer-wound coils, of this construction, require no end boards to hold the wires in place.

Coils without layer insulation (random-wound) may also be used, allowing about 50 per cent more turns in a given space. There are several disadvantages to this type of winding, however. The wire must have more than ordinary enamel insulation to prevent shorted turns, which reduces the space advantage somewhat. End boards, or tape, are necessary to hold the wires in place. Multiple winding cannot be used.

After the coil is wound, flexible lead wires are attached if the wire of the coil is smaller than 20 gage. Then the coil, or the core and coil together, are impregnated with a varnish. wax, or asphaltic compound to exclude moisture and air and to strengthen the coil mechanically. See Belden Handbook 12, Anaconda "Magnet Wire and Coils," or Inca Bulletin 3.

Inductors having a three-legged core, Fig. 1(*a*), (*b*), (*c*), have a single coil, placed on the middle leg of the core. With two-legged cores, Fig. 1(*d*), (*e*), (*f*), two coils are sometimes used, one on each leg, the two coils being connected in series or in parallel. The use of two coils results in lower resistance and a smaller dimension over the coil.

DESIGN PROCEDURE is carried out by:

(*a*) Choosing a core material and core size.

(*b*) Choosing a wire size, and determining how many turns of this wire will fit in the core window. About 5 per cent allowance should be made for wires not lying tightly together. Also, the total calculated build of the coil, including spool, layer insulation, and outside wrapper, should not exceed 90 per cent of the core window height.

(*c*) Calculating the inductance, resistance, core loss, heating, capacity, and Q from the dimensions and the number of turns. Three or four trial designs may be necessary before the desired constants are arrived at.

Inductance. Since sheet material cores are characterized by high permeability, practically all the magnetic flux is confined within the core structure. The inductance of a sheet-core inductor, without air gaps, is given by

$$L = \frac{4\pi N^2 A k \mu_{ac}}{l} \times 10^{-9} \quad \text{henry} \tag{1}$$

in which N is the number of turns on the coil, A is the cross-section of the core in square centimeters, k is the stacking factor, l is the mean length of the magnetic circuit in centimeters, and μ_{ac} is the a-c, or "incremental," permeability of the core material.

When air gaps are present in the core, the inductance is given by

$$L = \frac{4\pi N^2 A k \mu_{avg}}{l} \times 10^{-9} \quad \text{henry} \tag{2}$$

in which μ_{avg} is the average permeability of the core, including air gaps. This average permeability is

$$\mu_{avg} = \frac{\mu_{ac}}{1 + (a/l)\mu_{ac}} \tag{3}$$

in which a is the total effective length of all air gaps, in centimeters.

At any air gap the magnetic flux spreads, so that the cross-section of the magnetic field is greater than the cross-section of the core. It is most convenient to treat this "fringing" as if the length of the air gap were effectively reduced. If m and n are the dimensions of the core cross-section at the air gap, and l_g is the actual physical length of the air gap,

$$\text{Effective length} = \frac{mn}{(m + l_g)(n + l_g)} \times l_g \quad \text{approximately} \tag{4}$$

Permeability. Incremental, or a-c, permeability is the kind of permeability of interest in connection with most inductors. See *Magnetic Circuits and Transformers*, p. 198. It is a variable, depending upon:

(*a*) The material of the core.
(*b*) The amount of d-c magnetization.
(*c*) The amount of alternating flux.
(*d*) Wave form of the a-c voltage.
(*e*) Previous magnetization of the core.
(*f*) Temperature.

Of these factors, only the first three are considered in practical design work, although the others are by no means negligible.

When a core has no air gaps, the d-c magnetization, H_{dc}, is given by

$$H_{dc} = \frac{0.4\pi NI}{l} \quad \text{oersteds (or gilberts per cm.)} \tag{5}$$

in which I is the direct current flowing through the coil, in amperes.

It is usually more convenient to express the amount of a-c magnetization of the core in terms of flux-density variation, which is a function of the a-c voltage across the coil, rather than in terms of magnetizing force, which is a function of the a-c current in the coil. The peak a-c flux density, $B_{\max}$, is given by eq. (6), if the a-c voltage is sinusoidal.

$$B_{\max} = \frac{E}{4.44NAkf} \times 10^8 \quad \text{gausses} \tag{6}$$

In this equation, E is the rms voltage across the coil, f is the frequency, and the other symbols have the same meaning as given previously for eq. (1). Equation (6) applies whether or not there are air gaps in the core. See Figs. 4, 5, and 6.

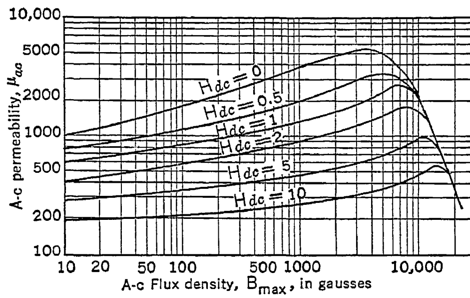


Fig. 4. A-c Permeability 3.6% Silicon Steel "58" Grade

D-c Magnetization with Air Gaps. D-c magnetization can be reduced by inserting an air gap in the core, improving the a-c permeability. See Figs. 4, 5, and 6. Up to a certain point, this results in an increase of inductance. See eqs. (2) and (3). Beyond that point, any further increase in the air gap causes a decrease of inductance.

The amount of d-c magnetization in the ferromagnetic core material, when there is an air gap in the core, can be determined by the graphical method shown in Fig. 7. See Karapetoff, *The Magnetic Circuit.* This method utilizes the normal magnetization or B-H curve of the particular core material, of which several are given in Fig. 8.

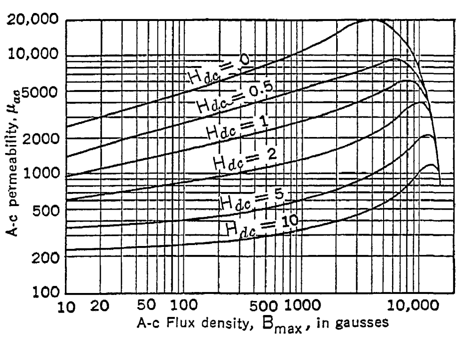


Fig. 5. A-c Permeability 50-50 Nickel-iron Alloys

The d-c magnetization of the core material and the a-c flux density [eq. (6)] having been determined, the a-c permeability is then found from curves such as Figs. 4, 5, or 6. This value of permeability is used in eq. (3) to compute the average permeability of the core, including air gaps, which is then used in eq. (2) to calculate the inductance of the reactor. It is usually necessary to try two or three values of air gap to discover the optimum one. This method of determining the optimum air gap, and the maximum inductance, may be called the "fundamental method."

A short-cut method of determining optimum air gap and maximum inductance was worked out by C. R. Hanna (Design of Reactances and Transformers Which Carry Direct Current, *Trans. A.I.E.E.*, February 1929). He showed that a curve can be drawn, for any particular core material, whose coordinates are NI/l and LI^2/V, V being the volume of the core. Such "optimum design curves" are shown for three commonly used materials in Figs. 9, 10, and 11. When using these curves it should be remembered that they apply only to a special case, in which maximum inductance for a given amount of direct current is the quality desired. If the reactor is to be used at several different values of direct current,

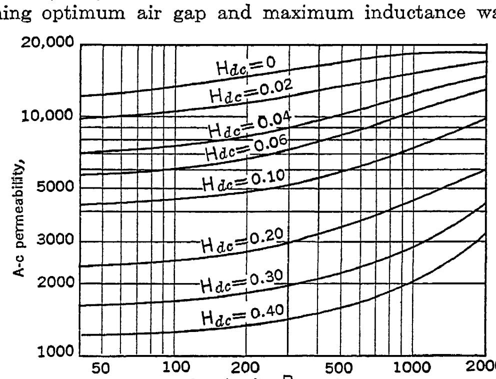


Fig. 6. A-c Permeability Mumetal

or if other considerations govern the design, the approach described previously should be used.

Saturation. If the sum of the d-c flux density and the peak a-c flux density approaches the saturation density of the core material, serious wave-form distortion occurs. This sum

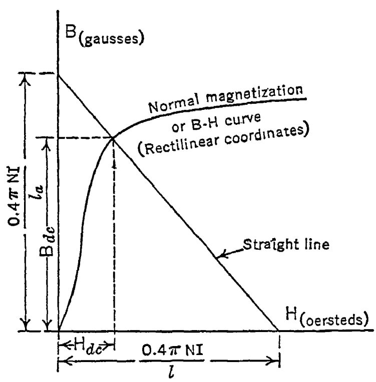


FIG. 7. Graphical Method of Determining D-c Magnetization of Core, when there is an airgap

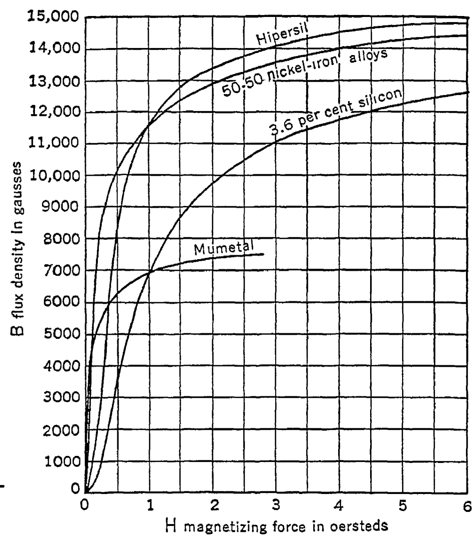


FIG. 8. D-c Magnetization Curves

should not exceed 12,000 gausses for silicon steel or for 50-50 nickel alloys. The d-c flux density may be found by the graphical method of Fig. 7, and the peak a-c flux density,

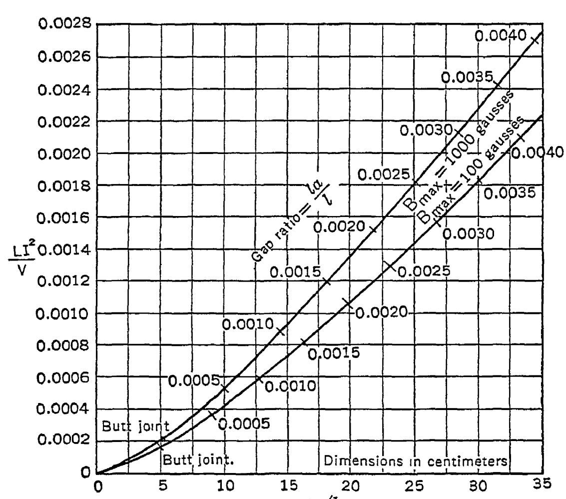


FIG. 9. Optimum Design Curves for Inductance with D.C. 3.6% Silicon

B_{max}, from eq. (6). If the total flux density is excessive, and the d-c flux density is the larger part, the air gap should be increased. If the a-c flux density is the larger part, the number of turns on the coil should be increased.

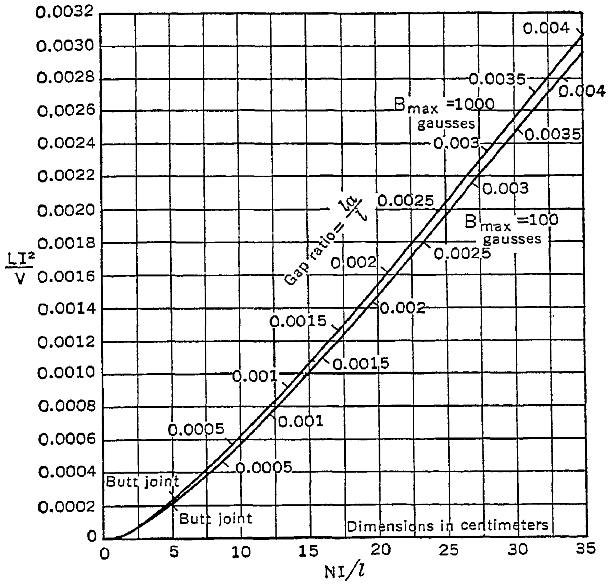


Fig. 10. Optimum Design Curves for Inductance with D.C. 50-50 Nickel-iron Alloys

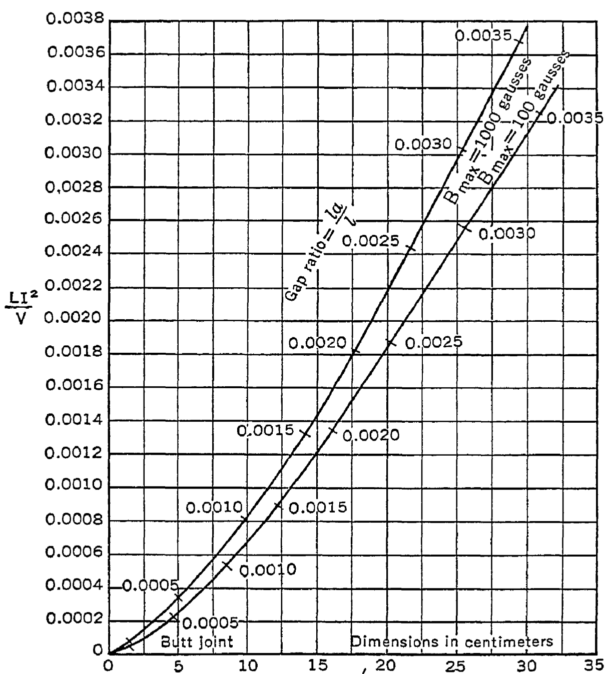


Fig. 11. Optimum Design Curves for Inductance with D.C. Hipersil

RECTIFIER-FILTER REACTORS AND AUDIOFREQUENCY PLATE REACTORS are examples of inductors for which maximum inductance at a given value of direct current is the most important quality desired. The "optimum design curves" are ideally suited for such designs.

THE SWINGING CHOKE is a reactor which must have a specified inductance at some large value of direct current, but which must increase rapidly in inductance as the direct current is decreased. An air gap is used which is optimum, not for the largest direct current nor for the smallest direct current, but rather for some intermediate value. An approximate design can be arrived at by using the "optimum design curves" for an intermediate value of direct current. Then the inductance should be calculated for the maximum and minimum currents by the fundamental method. In some cases, part of the core stack may have a large air gap to provide a good inductance at the largest current, while the remainder of the core may have a small air gap to give high inductance at the minimum current.

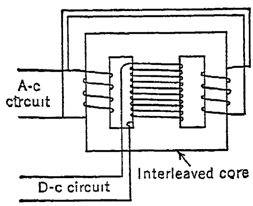


FIG. 12. Construction of Saturable Reactor

SATURABLE REACTORS are inductors the inductance of which is varied by means of a variable direct current. Interleaved laminations are used to get the greatest effect of the d-c magnetization. It is necessary to isolate the d-c circuit from the a-c circuit by means of separate windings. It is also necessary to prevent transformer coupling between the a-c and d-c windings. Figure 12 illustrates the construction of a saturable reactor. The a-c windings should always be connected in parallel. See Holubow, "D-C Saturable Reactors for Control Purposes," *Electronic Industries*, March 1945.

COIL RESISTANCE. The mean length of turn is calculated from the coil geometry, converted into feet, and multiplied by the number of turns. Reference to a wire table (Section 2, article 3) will give the d-c resistance of the coil, in ohms. Skin effect can be neglected, except for very large wire, above 1 kc. Allowance for temperature must be made. The resistance of a coil increases about 0.4 per cent per degree centigrade, above 20 deg cent. Thus, if an inductor operates at a coil temperature of 70 deg cent, its resistance will be 20 per cent higher than calculated for 20 deg cent.

CORE-LOSS curves are shown in Section 6, article 14. The core loss is determined in watts, which can then be added to copper loss watts to find the total heat dissipated in the reactor, and the temperature rise.

When the effect of core loss on the Q of the inductor, rather than heating, is the matter of interest, core loss can be represented by a resistor in parallel with the coil, of a value R_i ohms.

$$R_i = \frac{E^2}{W} \tag{7}$$

In this formula, E is the rms voltage across the coil, and W is the watts core loss. The core loss of most sheet-core materials varies very nearly as the square of the applied voltage, so that the ratio E^2/W is very nearly constant over a considerable range of voltage. Figure 13 gives core-loss curves of several materials at 2000 gausses. If P is the watts per pound, at 2000 gausses, and M is the weight of the core, in pounds, the core-loss resistance is

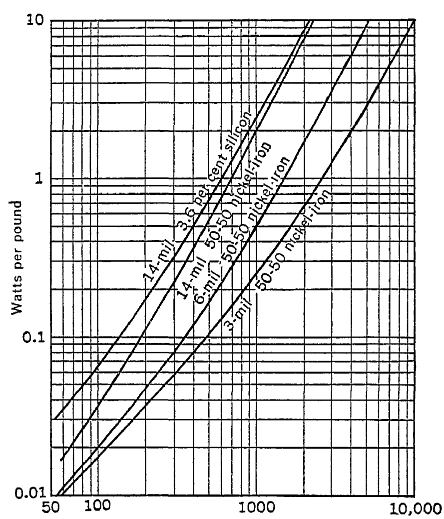


FIG. 13 Core Loss at 2000 Gausses (B_{max})

$$R_i = \frac{78.8(NAkf)^2}{PM} \times 10^{-10} \quad \text{ohm} \tag{8}$$

OPTIMUM Q. Q is often defined as the ratio of reactance to resistance. This is true when the only resistance present is in series with the coil, as is the case with the copper resistance of the winding. When there are both series and shunt resistances, as with a ferrous-cored inductor, Q is more clearly defined as the ratio of reactive volt-amperes to resistive volt-amperes.

For a particular core and coil, at a particular frequency, the shunt resistance of the core and the series resistance of the coil are practically fixed quantities. The reactance, however, can be varied by changing the air gap. There is a value of reactance which will produce the maximum Q. This problem is illustrated by the circuit shown in Fig. 14. In this circuit, R_c is the resistance of the coil, R_i is the resistance of the core, and X is the reactance of the inductor. Maximum Q is attained when

$$\frac{X}{R_c} = \frac{R_i}{X} \quad \text{or} \quad X = \sqrt{R_c R_i} \tag{9}$$

The air gap is adjusted so that

$$L = \frac{\sqrt{R_c R_i}}{2\pi f} \quad \text{for maximum } Q \tag{10}$$

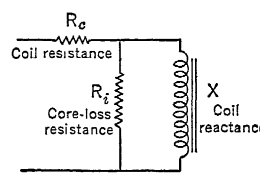


FIG. 14. Equivalent Circuit, Ferrous-cored Inductor

Equations (2) and (3) are used to compute the correct air gap. When the air gap and inductance are adjusted to optimum,

$$Q = \frac{X}{2R_c} \tag{11}$$

The maximum Q that can be obtained in a low-frequency, sheet-core inductor is practically independent of the inductance, the number of turns, the voltage, or the flux density. That is, a 10-henry reactor can be made with practically the same Q as a 1-mh reactor, by proper design. Maximum Q that can be realized depends only on the core material, the size of the core, and the frequency. See Fig. 15. For other sizes of core, the Q will

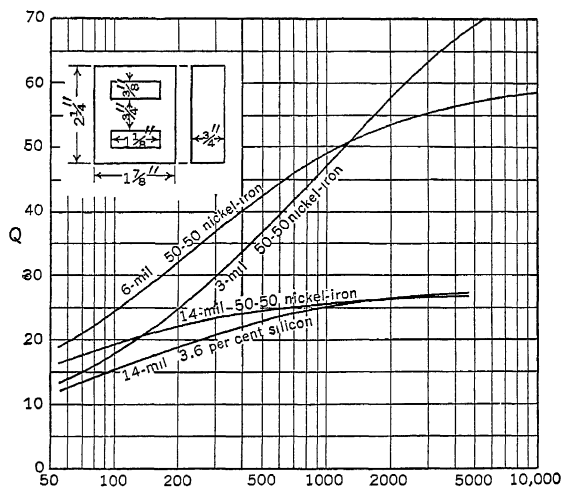


FIG. 15. Q vs Frequency, Various Core Materials

vary as the linear dimension if the proportions are the same, or as the cube root of the core volume if the proportions are different. Thus, a core having twice the linear dimensions of the one shown will have 8 times the volume but will give only twice the Q. For other core materials, the Q that can be realized will vary inversely as the square root of the relative core loss. Thus, a material having one-fourth the core loss will give twice the Q.

Air Gaps for Q should be located inside the coil, and preferably at the center of the coil length. There should be no air gaps outside of the coil. Fringing flux creates a magnetic field external to the reactor. When the reactor is placed in a metal can, or near a metal chassis, eddy currents are set up, which may reduce the Q of the reactor by a factor of 2 or 3 to 1.

Optimum Permeability for Q. For a particular core and coil, at a particular frequency, there is an optimum value of inductance which will give the highest Q. See eq. (10). This value of inductance requires a certain average permeability of the core. See eq. (2).

Practically all sheet-core materials have *too high* a value of a-c permeability to give the optimum inductance, so that it is necessary to *reduce* the overall permeability of the core by inserting an air gap. The use of an air gap to produce maximum Q in an inductor should not be confused with the use of an air gap to reduce d-c magnetization. The two purposes are different and distinct.

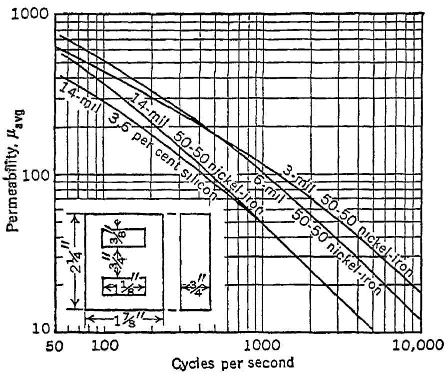


Fig. 16. Permeability for Maximum Q

The optimum overall permeability, μ_{avg}, required to produce maximum Q is a function of the core material, the core size, and the frequency. See Fig. 16. For other sizes of core, optimum permeability will vary inversely as the linear dimensions if the proportions are the same, or inversely as the cube root of the core volume if the proportions are different. For other core materials, the optimum permeability will vary inversely as the square root of the relative core loss.

A striking fact brought out by Fig. 16 is that the optimum overall permeability becomes less than 50 at frequencies above 4000 cycles. The high a-c permeability characteristic of sheet-core materials is of no advantage in this region. Permeability of the order of 50 can be realized with powdered-iron cores. This opens up the possibility of using powder cores at such frequencies.

16. HIGH-FREQUENCY, POWDERED-CORE INDUCTORS

Powdered cores differ from sheet cores in two important characteristics. They have, in general, lower core loss, particularly eddy-current loss, and they have much lower permeability. Both these characteristics are due to the subdivision of the core into minute particles, and both are desirable at high frequencies if the Q of the inductor is a major consideration. See above, "Optimum Permeability," and Fig. 16.

Material used for powdered cores are magnetite, a natural iron oxide, electrolytic iron, hydrogen-reduced iron, carbonyl iron, powdered Permalloy, and powdered molybdenum Permalloy. For descriptions of these materials and their method of manufacture, see Section 2, "Magnetic Materials"; H. G. Shea, "Magnetic Powders," *Electronic Industries*, August 1945; V. E. Legg and F. J. Given, "Powdered Molybdenum Permalloy," *B.S.T.J.*, July 1940.

Particle sizes range from about 50 microns diameter (0.002 in.) for audio-frequency cores to about 3 microns for cores useful at 100 mc. The particles making up a core are not all the same size. This is an advantage, as it gives a better packing of the magnetic material, the smaller particles filling in the spaces between the larger ones. Since eddy-current loss is proportional to the square of the particle diameter, the root-mean-square diameter of the particles making up a core is the particle dimension of greatest significance. Weight-average diameter is often given. This is a figure a few per cent higher than the root-mean-square diameter.

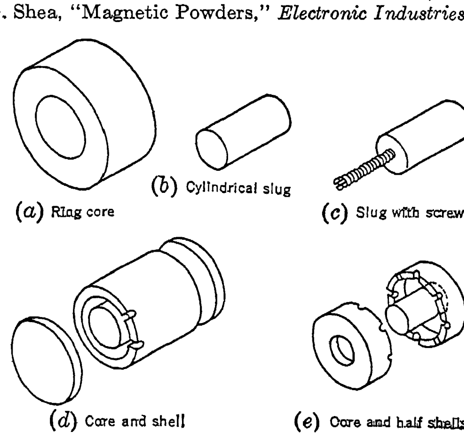


Fig. 17. Basic Types of Powdered-iron Cores

CORE CONSTRUCTION. Particles are coated with an insulating material, or treated in some manner to give them high surface resistivity, and mixed with a plastic binder. The mixture is compressed in molds at pressures of 50 to 100 tons per square inch, to form the desired core shape. The resulting core is a solid block of material resembling iron in appearance and somewhat lighter than iron in weight. Figure 17 illustrates a few basic styles. As better audiofrequency core materials are developed, cores of the shapes shown in Fig. 1 are also being offered.

COIL CONSTRUCTION, in general, follows the practice used for air-core coils of the same frequency range. Since powdered-iron cores are employed at the higher audiofrequencies and above, skin effect, coil capacitance, and coil dielectric loss are controlling factors in coil design, as with air-core coils. See Section 2, "Inductors with Air Cores." The presence of a core introduces two modifications of coil design. A somewhat longer coil, of less build-up, is desirable, to take better advantage of the core. Dielectric loss and dielectric constant of the winding form are of more importance when there is a core, especially if the core is grounded.

PERMEABILITY. "Intrinsic permeability" is the permeability of the magnetic particles. "Composite permeability," "ring-core permeability," or just "permeability" are terms used to describe the permeability of a core, made in the form of a ring sample, and wound with a toroidal coil. Such a sample has uniform flux density throughout its length and practically throughout its cross-section, and no end effects. Consequently, the ring-core permeability may be called the true permeability of the core material. The relation between ring-core permeability, μ, and intrinsic permeability μ_i, is given by

$$\mu = (\mu_i)^p \tag{12}$$

in which p is the "packing factor," or fraction of the core volume occupied by magnetic material. See Legg and Given, also analysis by H. Beller and G. O. Altmann, "Radio-Frequency Cores of High Permeability," *Electronic Industries*, November 1945. Table 3 gives values of ring-core permeability for a number of core materials.

The term "effective permeability" is used to describe the permeability of a cylindrical or slug-type core [Fig. 17(*b*)]. Such permeability is much lower than ring-core permeability of the same core material, because of demagnetizing effects at the ends of the cylinder. This matter has been investigated by R. M. Bozorth and D. M. Chapin, "Demagnetizing Factors of Rods," *J. Applied Phys.*, May 1942, and by W. J. Polydoroff and A. J. Klapperich, "Permeability of High Frequency Iron Cores," *Radio*, November 1945. Table 2 shows effective permeability versus ring-core permeability for various ratios of core length to core diameter. The coil is the same length as the core, in these data.

Table 2. Effective Permeability of Cylindrical Cores

Ring-core Permeability	Ratio Length/Diameter: 1	2	4	6	8
	Effective Permeability				
5	2.20	3.40	4.1	4.4	4.8
10	2.98	4.57	6.8	8.2	8.8
15	3.47	5.65	8.7	10.8	11.7
20	3.76	6.3	9.9	12.4	13.7
25	3.94	6.6	10.6	13.4	14.8
30	4.06	6.75	11.0	14.0	15.6

INDUCTANCE. The inductance of a toroidal coil on a ring core is given by

$$L = \frac{4N^2}{d}[A' + A(\mu - 1)] \times 10^{-9} \quad \text{henry} \tag{13}$$

in which N is the number of turns on the coil; d is the mean diameter of the core, in centimeters; A' is the mean cross-section of the coil, at right angles to the flux path, in square centimeters; A is the cross-section of the core, in square centimeters; and μ is the ring-core permeability of the core.

The inductance of a coil on a cylindrical, or slug-type core, if the coil is the same length as the core, is

$$L = L_0\left[1 + \left(\frac{b_i}{b_0}\right)^2\left(\mu_e - 1\right)\right] \tag{14}$$

In this equation, L_0 is the inductance of the coil without iron, b_i is the radius of the iron core, b_0 is the mean radius of the coil, and μ_e is the effective permeability, as given in Table 2. If the core is longer than the coil, the effective permeability is increased to a value

$$\mu_e' = \mu_e \sqrt[3]{\frac{l_i}{l_0}} \tag{15}$$

l_i being the length of the core, and l_0 that of the coil. This higher value should be substituted for μ_e in eq. (14).

Formulas are not available which apply to the various shell-type cores. The variety of shapes, non-uniform cross-section of the core, and square corners in the magnetic path offered by the core make any accurate analysis very difficult. The inductance of a coil, having a central core and a shell, will be greater than its inductance with the central core alone. On the other hand, its inductance will be less than that of a toroidal coil, of the same number of turns, on a ring core of the same average cross-section and the same mean length of magnetic path. See "Measurement of Iron Cores at Radio Frequencies," D. E. Foster and A. E. Newlon, *Proc. I.R.E.*, May 1941, and above reference by Polydoroff and Klapperich.

CORE LOSS AND Q. Core loss is of interest in high-frequency inductors primarily because of its effect upon Q, or quality factor. The addition of the core increases not only inductance but also resistance. A toroidal coil on a ring core has been analyzed by V. E. Legg, "Magnetic Measurements at Low Flux Densities," *B.S.T.J.*, January 1936. His formula for the increase of resistance due to core loss is

$$R = [(aB_m + c)f + ef^2]\mu L \quad \text{ohms} \tag{16}$$

In this, R is the resistance added to a coil by the core; a, c, and e are the hysteresis, residual, and eddy-current loss coefficients; f is the frequency, in cycles per second; B_m is the peak a-c flux density, in gausses; μ is the ring-core permeability; and L is the inductance of the coil with the core, in henries. Coefficients are given in Table 3 for a number of core materials.

The resistance added to a coil by a cylindrical core has been analyzed by Foster and Newlon, reference above, who give a formula

$$R = L_0 \frac{4\pi\mu_e^2 b_i^2 l_i \rho}{b_0^2 \sqrt{4b_0^2 + l_0^2}} \quad \text{ohms} \tag{17}$$

Symbols have the same meaning as for eqs. (14) and (15), ρ being the loss factor of the core material. The loss factor must be determined by measurements upon a sample of the particular core material. Published data are not available on the loss factor of cylindrical cores.

Table 3. Ring Permeability and Core-loss Coefficients

Material	Permeability	Hysteresis $\mu a \times 10^3$	Residual $\mu c \times 10^3$	Eddy Current $\mu e \times 10^6$
2-81 Molyb. Perm.....	125	0.20	3.8	2.4
2-81 Molyb. Perm.....	26	0.18	2.5	0.2
2-81 Molyb. Perm.....	14	0.16	2.0	0.1
81 Permalloy.........	75	0.41	2.8	3.8
81 Permalloy.........	26	0.30	2.8	0.7
Carbonyl "55".......	55	0.86	18.0	0.073
Carbonyl L...	39	1.45	27.0	0.10
Carbonyl L..........	24.8	3.1		0.13
Carbonyl C..........	16.7	1.1		0.14
Carbonyl E..........	10.4	0.3		0.11
Carbonyl TH........	9.6	0.3		0.10
Carbonyl SF.........	8.1	0.3		0.10
Electrolytic..........	23.4	2.4		0.33
Hydrogen-reduced...	42	1.04	23.0	0.17
Hydrogen-reduced....	18.4	2.6		0.12
Hydrogen-reduced....	16.9	1.0		0.12
Hydrogen-reduced....	12.5	3.1		0.11
Magnetite......	7.9	9.1		11.5
Magnetite...........	5.7	6.8		0.21
Magnetite...........	3.1	0.3		0.085

In Table 3, the figures for the permalloys are taken from Legg and Given. Figures for the other materials are General Aniline and Film Corporation data, published in the articles by H. G. Shea, and by Beller and Altmann, previously referred to.

PERMEABILITY TUNING, or "variable reluctance tuning," is the system of adjusting a circuit to resonance, at a desired frequency, by moving an iron core in or out of the coil. A simple arrangement for doing this is to mold a screw in one end of a cylindrical core as illustrated in Fig. 17(c). The screw passes through a tapped hole in the coil housing and is slotted at the end for a screwdriver. Rotation of the screw moves the core axially into or out of the coil, varying its inductance and the resonant frequency of the circuit. Such a system is ideal for a circuit tuned to some fixed frequency, such as an intermediate-frequency transformer circuit. See "Ferro-inductors and Permeability Tuning," W. J. Polydoroff, *Proc. I.R.E.*, May 1933.

Incremental permeability tuning is a system of adjusting the resonant frequency of a circuit by varying the permeability of the iron core without any mechanical motion. The permeability is varied by means of d-c magnetization on the same principle as the saturable reactor. Increase of d-c magnetization causes a decrease of inductance and an increase of frequency. By proper design, the increase of frequency, from some minimum value, may be made proportional to the direct current. "Incremental Permeability Tuning," W. J. Polydoroff, *Radio*, October 1944.

CAPACITORS

By James I. Cornell

A capacitor or condenser is an electrical device used primarily because it possesses the property of capacitance. Though *capacitor* is the preferred engineering terminology, the term *condenser* is still widely used.

Electrostatic capacitance is defined as the ratio of the electrical charge Q stored in the capacitor by virtue of the applied voltage (E_{dc}). That is, when a d-c voltage is impressed on two conductors insulated from each other, the voltage causes an electrical charge to flow into the system. One conductor assumes a positive charge and the other an equal negative charge, depending upon the polarity of the impressed voltage. The charge on the conductors produces electrostatic stresses in the region between them. The work done in charging the capacitor appears as stored potential energy. This energy is released when we remove the impressed voltage and short-circuit the capacitor electrodes. The capacitance is a measure of the charge or stored potential energy for a given voltage. In terms of physical units, a capacitor having a capacitance of 1 farad will store 1 coulomb of charge or 1 watt-second of energy for 1 volt of applied direct voltage.

17. CLASSIFICATION OF CAPACITORS

Capacitors may be classified according to *form*, *dielectric medium*, and *electrode structure*.

Generally, the classification of capacitors according to form is determined by whether the capacitance is *variable* or *fixed*.

Variable capacitors cover two essential types, namely:

1. *Variable capacitors* provide for continuous control of the capacitance and are used for varying the resonance frequency of a tuned circuit with which it is associated. The dielectric employed may be air, compressed gas, or liquid types. See Fig. 1.

2. *Adjustable capacitors* provide for limited control of the capacitance, and these capacitors are usually found in frequency-determining circuits for alignment purposes such as intermediate-frequency transformers, etc. The dielectric medium used in adjustable-type capacitors may be air, mica, or some form of ceramic.

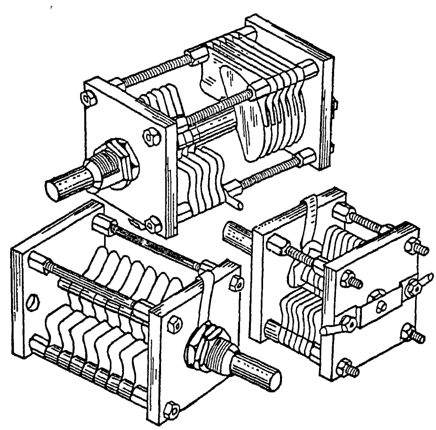


FIG. 1. Typical Variable Capacitors

Fixed capacitors employ a wide range of dielectric materials, and, because of the different dielectric media used, the form and usage of these capacitors must be evaluated according to their dielectric in terms of circuit requirements. The wide variation in the kinds of dielectric materials available for capacitors permits construction to meet practically all kinds of circuit requirements.

FIXED CAPACITORS CLASSIFIED ACCORDING TO DIELECTRIC MEDIUM.

Gas Dielectrics. Vacuum and compressed-gas-filled capacitors are designed for electronic power circuits where high voltage, high current, and high frequency requirements are encountered and where other types of dielectrics are inadequate because of excessive losses and resultant overheating.

Impregnated Paper. The most common form of solid-dielectric capacitors employed in electronic and communication circuits are impregnated-paper types. High-purity kraft paper may be impregnated with microcrystalline hydrocarbon waxes, chlorinated waxes, vegetable oil, mineral oil, Askarels or synthetic chlorinated oils, and plastics. Impregnated-paper capacitors are usually made up in a multiple-layer foil and paper structure of rolled construction and are impregnated after winding.

Plastic Film. Special design considerations such as low dielectric absorption or high operating temperatures have resulted in the development of plastic-film dielectrics, using such substances as polystyrene, acetate, or butyrate. These follow the general construction of the impregnated-paper types except that the plastic dielectrics are in final form and are impregnated only for the purpose of removing surface moisture and eliminating voids in the winding.

The principal virtue of polystyrene lies in the very low absorption characteristic and its low losses at radio frequencies. Its use is limited to an operating temperature of 85 deg cent. Other synthetic plastic films such as acetates and butyrates have been found useful for applications requiring operation at ambient temperatures exceeding 85 deg cent, and they have been operated experimentally at 150 deg cent. They usually have high r-f power factor, and their use is restricted to d-c or low voltage a-c circuits where high ambient temperature is the principal consideration.

Inorganic Dielectrics. Common inorganic dielectric materials are solid dielectrics consisting of mica, glass, and ceramic types; also liquid dielectrics such as silicone oils. The solid dielectrics are usually in sheet or plate form which require a laminated stacked construction. The electrodes may take the form of foil, or a silver coating deposited directly on the surface of the dielectric before the stacking operation.

Ceramic dielectrics are supplied in a variety of shapes other than laminal, and one of the most popular is a hollow cylinder with the electrodes placed on the inner and outer surfaces in the form of a silver coating. They form a capacitor whose capacitance is a function of the dielectric constant of the ceramic. Capacitance values are usually low or less than 2000 $\mu\mu$f. The dielectric constant and temperature coefficient of the ceramic body can be varied widely to give capacitors with negative, positive, or zero temperature coefficients of capacitance.

This form of capacitor is receiving more consideration in r-f circuits where a negative temperature coefficient is used to compensate for the positive temperature coefficients of other circuit elements with which they are associated.

The electrode structure used with a silicone oil dielectric is of the rigid grid type similar to that found in variable capacitors. Silicone oils will withstand high ambient temperatures and are characterized by low r-f power factor.

Oxide Film Dielectrics. Electrolytic capacitors owe their unusual characteristics to the oxide film dielectric layer which is produced electrochemically on aluminum, tantalum, and certain other metals. Details of this form of dielectric will be found under the subject "Electrolytic Capacitors."

CLASSIFICATION OF FIXED CAPACITORS ACCORDING TO PLATE STRUCTURE. Rigid multiple parallel plate structure used in connection with gas or liquid dielectrics. Under this classification may be found the silicone and compressed gas capacitors for high-power, high-frequency radio transmitting and dielectric heating circuits; also wet electrolytic capacitors for use in low-voltage receiver power-supply filters.

Interleaved foil or stack construction used in connection with mica, glass, ceramic, and plastic film solid dielectrics.

Helical or Rolled Plate Construction. The most common form of plate structure for solid dielectrics such as impregnated paper, plastic film, and electrolytic types is the helical or rolled plate construction. In this construction the electrodes are of very thin foil separated by a single- or multiple-ply dielectric layer and wound into a cylindrical roll.

Metallized plate construction where the electrodes are deposited on the surface of the dielectric. This construction is used with mica, ceramic, glass, plastic, and impregnated-paper dielectrics.

CAPACITOR CHARACTERISTICS. Variable Nature of Capacitance. Three principal characteristics of any dielectric medium influence the physical form of capacitors:

1. Dielectric constant or specific inductive capacity, denoted by K.
2. Dielectric strength.
3. Dielectric loss or power factor.

The dielectric constant is a numerical quantity, expressed as the ratio of the capacitance of the structure with a dielectric other than air to the capacitance with air as the dielectric. In other words, the higher the specific inductive capacity or dielectric constant (K), the smaller the size for a given capacitance. *Note:* Refer to Section 2 for the dielectric constant of various materials.

Dielectric strength may be defined as the property of the dielectric by which it withstands breakdown when a voltage is applied to it. The dielectric strength is expressed in volts per mil of dielectric thickness. The shape of the electrodes by which the voltage is impressed on the dielectric and the duration of the impressed voltage help to determine the rupture voltage. When the applied voltage is alternating, the wave shape and frequency of the voltage affect the dielectric strength. Because of these variables, the dielectric strength in volts per mil is not a constant and is usually expressed as a voltage range.

Capacitance varies inversely as the plate separation or dielectric thickness. The limiting minimum thickness is the voltage that the capacitor is required to withstand. Therefore, the higher the dielectric strength, or breakdown potential, of the insulating medium, the thinner the dielectric and the larger the capacitance value for a given electrode area. This can best be illustrated by comparing ceramic dielectric to the oxide film dielectrics of electrolytic capacitors. Most ceramic dielectrics are limited to a breakdown potential of less than 100 volts per mil, whereas the oxide film thickness of an electrolytic capacitor is measured in microns and will withstand a voltage stress equivalent to a million volts per mil of dielectric thickness.

D-c Leakage. Continued polarization of a capacitor by direct voltage after full electrification results in a flow of current termed the leakage current. Though the d-c leakage is negligible under most conditions of operation, it varies with temperature and should be taken into account in circuit designs, especially those involving grid coupling capacitors or similar circuits which are affected by the flow of conduction current in the capacitor.

Dielectric Polarization and Absorption. Capacitors with solid dielectrics take longer to charge than would be predicted from their theoretical constants, owing to a lag or delay in polarizing the dielectric medium. This is known as dielectric polarization. When a capacitor is discharged by short circuiting, all its energy is not released and it will build up a new charge with time which is known as the residual effect. This characteristic is known as dielectric absorption. Both are detrimental to circuits requiring rapid charge and discharge characteristics. Selection of a dielectric like polystyrene is dictated in circuits where dielectric absorption must be held to the absolute minimum.

Figure 2 illustrates schematically what takes place in a capacitor having d-c leakage and dielectric absorption. C represents the geometric capacitance based on a perfect dielectric. C_1 and r_1 represent the absorption effect. The pure conduction effect is represented by r_2.

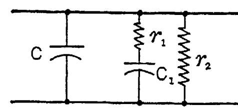


FIG. 2. Schematic Circuit of Capacitor

Power Factor, Q, or A-c Resistance. When a capacitor is operated under a-c voltage stresses, the dielectric will heat up, depending upon the frequency of the applied voltage. The a-c resistance or the power factor is a measure of the heat dissipated. It is more convenient in working with electrical circuits to use a ratio of the pure reactance to the effective resistance which is termed the Q factor. The Q factor can be employed for the purpose of comparing various capacitors quantitatively since it constitutes a figure of merit for a given design.

It is easily seen that, although so-called fixed capacitors may be made up using any one of the many different types of dielectrics, the actual capacitance is a variable depending upon the physical characteristics of the dielectric, which vary with:

A. *Voltage.* Capacitance will change with voltage. The voltage coefficient will vary with the magnitude of the d-c or a-c impressed voltages.

B. *Frequency.* The capacitance will change with frequency and is a function of the Q or power factor of the dielectric.

C. *Time.* Capacitance will change with time because of dielectric aging.

D. *Temperature.* Capacitance values will change with temperature since dielectrics change with temperature or have a temperature coefficient.

These characteristics serve to demonstrate the fallacy of considering a capacitor as an ideal fixed capacitance.

Inductance. In addition to the limitations imposed by the dielectric media, another limitation, which results from mechanical design, is inductance. All capacitors have a self-resonant frequency and behave like a circuit involving series inductance, resistance, and capacitance. Above the critical frequency, the reactance is inductive and not capacitive. In fact, capacitors behave like a complex impedance depending upon the operating frequency range.

18. VARIABLE AND ADJUSTABLE CAPACITORS

Capacitors designed for frequent adjustment by an equipment operator are usually termed variable or tuning capacitors. Capacitors of this type most often use an air dielectric, although other dielectrics are suitable for special circumstances. For example, compressed air or nitrogen dielectrics reduce the size of high-voltage capacitors in large transmitters. Mineral oil, silicone fluid, and other liquid dielectrics save space because of their higher dielectric constant. In addition, such capacitors have a higher breakdown voltage for the same interelectrode spacing.

Capacitance adjustment in variable capacitors is usually made by varying the effective plate area. The capacitor consists of two sets of parallel intermeshed plates, one fixed and one rotatable on a mounting shaft. Rigidity of the capacitor framework and freedom from warpage of the plates and the stator insulation are extremely important from the standpoint of circuit stability. For extremely stable circuits in high-grade electronic equipment, carefully designed Invar steel frames and plates may be used to minimize capacitance shifts with ambient temperatures. More commonly frames and plates are of aluminum or silver-plated brass. In the cheapest broadcast receivers, cadmium-plated steel has been used. To facilitate tracking of circuits in broadcast receivers, the outer rotor plates are sometimes slotted to permit small adjustments in the capacitance-rotation curve. The supporting insulation for the capacitor stator is usually phenolic or ceramic, depending upon circuit considerations. Electrically low-loss, dimensionally stable steatite or glass-bonded mica insulation is used in capacitors where high Q and low capacitance drift are important. In the highest grade of precision laboratory standard capacitors the insulation is quartz.

In certain highly accurate instruments, the "standard" variable capacitors consist of two co-axial cylinders, one fixed and one movable.

Capacitors intended for relatively infrequent adjustment of capacitance are termed "trimmer or adjustable" capacitors. Most common in broadcast receivers is the small mica trimmer capacitor. Such a capacitor consists of a fixed and a hinged movable metal electrode mounted on a phenolic or ceramic base with a mica spacer between the two electrodes. The movable leaf is raised or lowered by threading it on a screw. These capacitors are sometimes called "book mica trimmers."

In high-grade electronic equipment, especially for frequencies above the standard broadcast band, trimmer capacitors similar in construction to intermeshed plate variable capacitors are used. In many cases, they have a shaft positioning locking device such as a split tapered bushing and nut. Also found in such equipment are adjustable ceramic capacitors. These capacitors consist of two coaxial half-silvered ceramic disks, one fixed and one rotable. The ground unsilvered faces are kept in contact by spring pressure. The capacitance depends on the extent of overlap of the silvered faces. Through selection of various ceramic compositions an opportunity is provided for some degree of circuit temperature compensation.

Transmitter neutralizing capacitors are a special design of adjustable capacitors consisting of a fixed disk with rounded edges and a similar coaxial movable disk mounted on the end of a set screw. Capacitance adjustment is by turning the screw in and out of a threaded support.

In high-powered aircraft transmitters, adjustable vacuum capacitors are sometimes employed. These capacitors have a metal bellows to permit positioning the movable capacitor element.

The maximum capacitance of variable capacitors usually employed in electronic circuits is from 10 to 530 $\mu\mu$f. For use as a capacitance standard, units are made with capacitances up to 5000 $\mu\mu$f. Trimmer capacitors may have a maximum capacitance of 5 to 75 $\mu\mu$f.

Since air dielectric capacitors have very little loss in the dielectric and supporting insulation, they will show practically no change in capacitance with frequency. Accurate measurements of capacitance values at audio frequencies may therefore be depended upon at radio frequencies with a good degree of precision.

EFFECT OF STRAY CAPACITANCE. Units of variable capacitance generally are for relatively small values, and in a circuit, particularly at high frequencies, the capacitance effects of the various other parts of a circuit will be appreciable as compared with the capacitor unit. Every part of the apparatus has capacitance to other parts, and these small stray capacitances may be appreciable. The stray capacitances are particularly objectionable because they vary when parts of the circuit or conductors near by are moved, such as the hand or body of the operator. The disturbing effects may be minimized in practice as follows: (1) by keeping the capacitor a considerable distance away from conducting or dielectric masses; (2) by shielding the capacitor, that is, surrounding the whole capacitor by a metal covering connected to one of the sets of plates; (3) by using a capacitor of sufficiently large capacitance so that stray capacitances are negligible in comparison. The first two of these methods reduce only the stray capacitances of the capacitor itself to other parts of the circuit and to external moving bodies. Although the third procedure is workable for the lower radio frequencies, it fails at the very high and ultra-high frequencies, and an entirely different approach to the problem has been necessary.

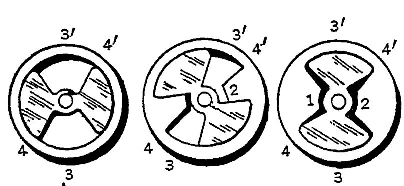


Fig. 3. Semi-butterfly Circuit

Karplus has described a variable capacitor design which includes the circuit inductance. In this design, wide tuning range with simultaneous change of lumped capacitance and inductance is obtained by rotation of a member that does not require any electrical connections. This arrangement is called the butterfly or semi-butterfly circuit, depending upon the configuration of the variable capacitor design. Figure 3 shows such a circuit.

Maximum and minimum capacitance of a butterfly circuit can be computed as in a variable air capacitor, but, owing to the butterfly shape, capacitance ratios are considerably less than in well-designed tuning capacitors.

CAPACITANCE FORMULA. The general expression for the capacitance of a multiplate capacitor is (in micro-microfarads)

$$C = 0.0885K\frac{(N-1)S}{T} \tag{1}$$

in which S is the area in square centimeters of a moving plate overlapping a fixed plate; T is the separation of plates in centimeters; K is the dielectric constant (for the air capacitor, $K = 1$); N is the total number of similar plates (fixed plus movable), alternate plates being connected in parallel. In eq. (1) no correction is made for the curving of the lines of force at the edges of the plates; this effect is negligible for most capacitors, when T is very small compared with S.

The numerous applications of variable capacitors in radio circuits place different requirements upon the capacitor in its characteristic variation with setting. For laboratory use a capacitor is usually designed with semicircular plates. With this form the capacitance increases linearly with the angular displacement of the movable plates. In other radio-circuit applications it is convenient to have the capacitance vary with the setting to some other power than the first, and these forms are noted below.

Plate Form, Semicircular. In a variable capacitor of the semicircular type the effective area of the plates is changed by rotating the movable plates. Throughout the entire plate rotation the capacitance is proportional to the setting, provided that the capacitor is well constructed and the distance between the two sets of plates is not affected by rotation of the movable set. When the plates are entirely unmeshed (no overlap), measurement usually will show that there is an appreciable capacitance between the capacitor terminals. For a 5000-$\mu\mu$f capacitor this may be as large as 75 $\mu\mu$f. This value is practically unaffected by the position of the movable plates on the rotor, and it represents the capacitance between the insulated binding posts and the small capacitance formed across the insulators which separate the plates. The capacitance of a variable capacitor cannot, therefore, be reduced to zero, and such relations as are developed for capacitance-rotor positions hold only when the plates are enmeshed a considerable amount. Throughout the range in which the capacitance curve is linear the capacitance is given by the expression $C = a\theta + b$, in which θ is the angle of rotation of one set of plates with respect to the other.

Plate Form, Logarithmic. In a special form of wavemeter, called the decremeter, the plates are so formed as to determine the logarithmic decrement of a circuit. This is measured by the percentage change in capacitance required to reduce by a certain amount the indication of an instrument in the circuit at resonance. In order that equal angular rotations may correspond to the same decrement at any setting of the capacitor, it is necessary that the percentage change in capacitance for a given rotation shall be the same at all parts of the scale. In order that each degree displacement give the same percentage increase in capacitance the following requirement must be satisfied:

$$\frac{dC}{C} = a\,d\theta \tag{2}$$

where a = constant = percentage change of capacitance per scale division. Integrating,

$$\log C = a\theta + b \tag{3}$$

where b = a constant

$$C = e^{(a\theta+b)} = C_0e^{a\theta} \tag{4}$$

where $C_0 = e^b$ = capacitance when $\theta = 0$.

Since the area must vary as the capacitance:

$$A = {}^1\!/_2\int r^2\,d\theta = C_0e^{a\theta} \tag{5}$$

using polar coordinates

$$\frac{dA}{d\theta} = \frac{r^2}{2} = C_0ae^{a\theta} \tag{6}$$

and

$$r = \sqrt{2C_0ae^{a\theta}} \tag{7}$$

The last expression is the polar equation of the bounding curve required to give a uniform decrement scale.

Plate Form, Wavelength. For certain applications it is desirable to have the wavelengths proportional to the setting of the capacitor. Since the wavelength varies as the square root of the capacitance, this will require that the capacitor plates be formed so that the capacitance of the circuit varies with the square of the setting. In designing the shape of these plates, allowance should be made for the stray capacitance of the circuit. For the design of the simple straight-line wavelength capacitor the following expressions are used:

$$C = A = a\theta^2 \tag{8}$$

In polar coordinates the area

$$A = {}^1/_2 \int r^2 \, d\theta \tag{9}$$

Differentiating both expressions:

$$\frac{dA}{d\theta} = \frac{r^2}{2} = 2a\theta$$

$$r = \sqrt{4a\theta} \tag{10}$$

which is the polar equation of the bounding curve to give the desired characteristic.

Plate Form, Frequency. In another type of capacitor, of much more importance than those mentioned above, the variation of capacitor setting is directly proportional to the frequency for which the total circuit is tuning. The importance of this type follows from certain considerations in radio communication, which require the spacing of broadcasting stations from each other by equal increments in frequency. It is this condition which makes highly desirable a capacitor so designed that equal increments in capacitor setting advance by equal increments the frequency to which the associated circuit is tuned. Such a capacitor will tune for the various broadcasting stations at equally spaced points in the capacitor dial.

In making this special form of variable capacitor the shapes of either rotor or stator plates may be adapted to perform the required area-setting variation. It is usually more convenient to use ordinary semicircular stator plates and form the rotor plates to the required shape. On this basis Forbes has shown that the radius vector to the edge of the rotor plate must satisfy the relation:

$$r = \sqrt{\frac{4D^2}{nkK^2\left[\dfrac{D}{K\sqrt{C_0}} - \theta\right]^3} + r_1^2} \tag{11}$$

in which $D = 1/(2\pi\sqrt{L})$; L = inductance of the circuit in henries; n = number of dielectric spaces; $k = 10^{-11}/(36\pi d)$; d = length of air gap between plates; $K = (f_{0°} - f_{180°})/\pi$ = cycles per radian of capacitor scale; $f_{0°}$ = frequency of circuit with capacitor set at 0 deg; $f_{180°}$ = frequency of circuit with capacitor set at 180 deg; C_0 = total capacitance of circuit when capacitor is set at 0 deg—this includes stray circuit capacitance as well as that of the zero setting of the capacitor in farads; θ = angle of rotation of the plates, radians; and r_1 = radius of the cut-out of stator plates to accommodate the rotor shaft. All dimensions in centimeters. The capacitance of such a capacitor, for any angular position θ of the rotor, is

$$C_\theta = \left[\frac{D}{\dfrac{D}{\sqrt{C_0}} - K\theta}\right]^2 + \left[\frac{nkr_1^2}{2}\theta\right] - [C_0] \tag{12}$$

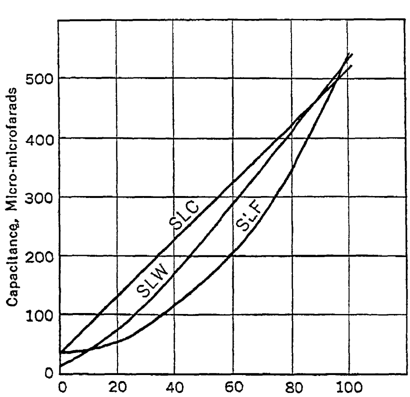


FIG. 4. Capacitance Calibrations for Three Typical Forms of Variable Air Capacitors, Straight-line Capacitance, Straight-line Wave Length, and Straight-line Frequency

Typical variations of capacitance with setting are shown in Fig. 4, which represents the construction practice of one manufacturer of variable capacitors. The curves shown indicate the value of capacitance obtained for various settings of the three general types of capacitors, straight-line capacitance (SLC), straight-line wavelength (SLW), and straight-line frequency (SLF).

Plate Form, Special Designs. Special considerations must be given to the design of variable air condensers when two or more units are connected together for operation on a common shaft. This grouping of sections into "gangs" sometimes places five condensers on a single control, and these may perform important functions in as many individual circuits. An illustration of this construction occurs in a radio receiver in which there are two identical units for tuning signal frequency circuits, one unit with specially shaped plates to tune the oscillator circuit and two smaller units to tune the short-wave circuits. In order to integrate the actions of all these condensers in the above superheterodyne radio receiver, where the single control is particularly desirable, various methods of design have been developed. E. D. Koepping has described the general principles involved, D. F. McNamee has presented a graphical solution of the design problem, and H. Schwartzmann has shown an analytical solution of the problem (see Bibliography, p. 3-73).

PLATE SPACING OF AIR CAPACITORS. Between two oppositely charged parallel plates of infinite extent the field is uniform. It is known, however, that the field between two finite plates is not uniform, tending to greater voltage gradients at the edges. Stress at a point in a dielectric is determined by the gradient at that point, and stresses above certain critical values lead to the formation of corona and sparkover. This may be pictured as follows, with plate spacing = S and thickness as T. For large values of S/T, the radius of the plate edge is small and a high gradient exists. As the plate thickness is increased, keeping plate center-to-center spacing and applied voltage constant, the radius of edge curvature is increased. This decreases the edge gradient. Inside the capacitor, however, the gradient increases because of the reduced spacing. It has been suggested by Ekstrand that the best value of S/T may be between 2 and 3. Investigation seems to corroborate Ekstrand figures by fixing the optimum S/T at 2.77.

Plate Dimensions and Voltages of Capacitors at Various Frequencies

T in.	S in.	S/T	Sparkover Voltage			Maximum Gradient		
			kv @ 60 cy.	kv @ 700 kc	kv @ 1500 kc	kv/in. @ 60 cy.	kv/in. @ 700 kc	kv/in. @ 1500 kc
0.128	0.218	1.705	14	13.5	13.7	89.1	85.8	87.1
0.04	0.192	4.8	8.4	7.59	6.82	80.2	72.5	75.1
0.064	0.719	11.24	24	14.28	11.7	82.0	48.8	40.0

It has become a habit to consider breakdown as coincidental with sparkover. Actually, breakdown has occurred at the first sign of corona, and corona may become evident at a considerably lower voltage than sparkover. When a conductor is raised beyond a certain critical potential, the air adjacent to it becomes ionized, forming corona. The ionized air,

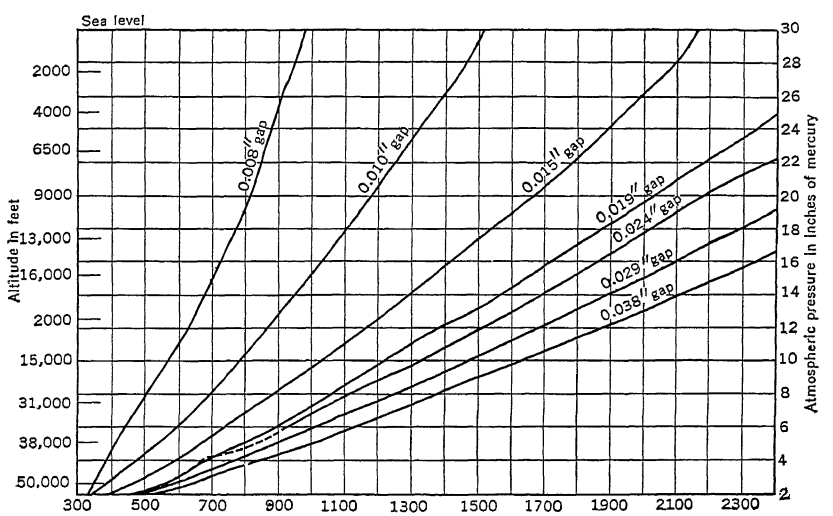


FIG. 5. Breakdown Voltage vs Altitude (Atmospheric Pressure)

being itself a conductor, can be considered to increase the dielectric losses. If the gradient is further increased, the conditions become unstable and sparkover occurs.

Not only are capacitors used in transmitters and electronic heating devices subjected to high voltages, but they also must carry considerable amounts of power. The power input to a capacitor is measured by the energy stored in it multiplied by the number of charges and discharges per second.

Transmitter capacitors used in aircraft electronic equipment must have their plates spaced properly to withstand voltages encountered at maximum flying altitudes. The curves of Fig. 5 indicate breakdown voltage for various plate spacings at various altitudes.

19. IMPREGNATED-PAPER CAPACITORS

Impregnated-paper capacitors are the most efficient of all types because of their flexibility in size, shape, and rating. They cover an extremely wide range in size from the small toothpick varieties found in hearing aids to large welded case blocks. Flexibility in voltage ratings is second only to size values with a range from 1 volt to 200,000 volts. Capacitance values are common in a range from $100\mu\mu f$ to $200\mu f$ in a single container.

Electronic equipment designers are primarily concerned with d-c application of impregnated-paper capacitors.

D-c service includes such applications as rectifier filters, energy storage, arc suppression, and by-passing for electron-tube and circuit elements. D-c ratings provide for small a-c components where their heating effects are negligible. Experience indicates that the a-c

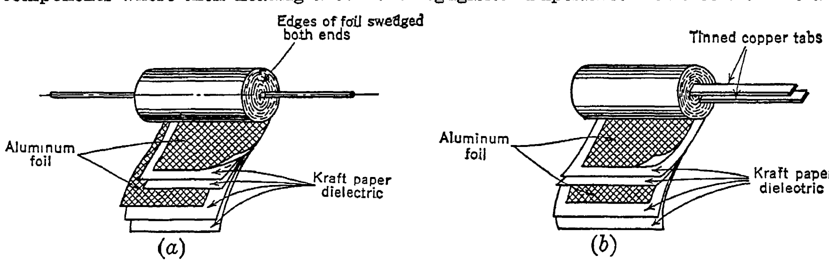


FIG. 6. Inductive and Non-inductive Capacitor Windings

components should not exceed 20 per cent of the d-c value at a frequency of 60 cycles, or 15 per cent for 120 cycles, or 1 per cent for 10,000 cycles, and for higher frequencies the allowable magnitude is determined strictly on thermal evaluation of the high frequency component.

Other special electronic applications require rating of the capacitor on an a-c rather than a d-c basis, such as tuned filters and pulse networks.

A-c versus D-c Ratings. It is possible to rate capacitors designed for a-c service in terms of d-c voltages, but it is not practical to rate d-c capacitors in terms of a-c voltage ratings because of the difference in design considerations in the d-c rated capacitors in comparison to those rated for alternating current.

Some of these design considerations are:

1. D-c voltage ratings depend essentially on dielectric stress.

2. A-c voltage ratings depend not only on dielectric stress but also on the operating frequency and power factor. Frequency and power factor determine the internal heating, which must be kept within limits determined by the radiating surface of the container. For small a-c capacitors, the ratings are determined from dielectric breakdown voltage considerations. In the larger a-c voltage ratings, the prime design consideration is heating.

In general, the following tabulation represents the nearest standard d-c rating corresponding to the several standard a-c voltage ratings:

A-c Volts	D-c Volts
110	200
220	400
330	600
440	1000
550	1500
660	2000

Chlorinated diphenyl impregnated capacitors designed for a-c service cannot be used on direct current unless the impregnant has been chemically treated with inhibitors or stabilizers to prevent deterioration of the impregnant from the combined influence of the d-c field and high temperature.

Construction of Winding. Impregnated-paper capacitors are made in a roll construction, consisting of two metallic foils separated by two or more sheets of impregnated kraft tissue. In the roll construction the resultant capacitance is twice that obtained with a parallel-plate construction since both sides of the foils are active. The capacitor winding may be round or flat, depending upon mechanical considerations of housing.

A further consideration in roll construction is whether the foils are of the "buried" type with tabs for contact members or of the extended foil construction. See Fig. 6(*a*) and (*b*). The extended foil construction gives the lowest value of self-inductance since all the turns are bonded together, the construction approaching that of a stacked parallel-plate capacitor. The "buried" foil winding approaches the "extended" foil if the tabs are inserted at the center of the winding within one turn of each other. Figure 7 shows a comparison of the impedance for "extended" vs. "buried" foil constructions for a frequency range of 60–5000 cycles.

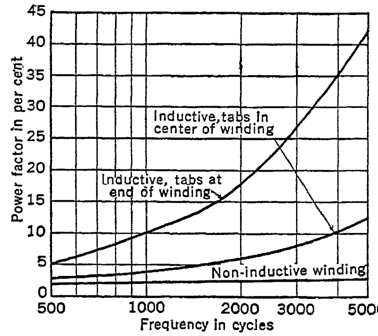


FIG. 7. Comparative Power Factors for Inductive and Non-inductive Windings

Impregnation of Winding. The impregnants most commonly used with kraft paper are microcrystalline hydrocarbon waxes, chlorinated waxes and oils, castor oil, mineral oil, and plastic compounds. These various impregnants offer a wide range of characteristics with temperature. Figure 8 shows six of the more common wax and oil impregnants for a temperature range of −40 to 100 deg cent. These data were taken for average production capacitors and do not represent minimum requirements. Minimum requirements are shown in the accompanying Table 1.

Voltage Rating of D-c Capacitors with Temperature and Service Conditions. All capacitors are affected by temperature, voltage stress, and time.

Table 1. Minimum Impregnant Requirements

	Castor Oil	Mineral Oil	Chlorinated Diphenyl	Halowax
Megohms times microfarads:				
At 25° C	500	2000	1500	2000
At high test temperature *	5	20	15	100
Insulation resistance in megohms:				
At 25° C	1500	6000	4500	6000
At high test temperature	150	600	450	1000
Capacitance change in per cent at low ambient test temperature from value at 25° C †	−30	−15	−30	−10

* 85° C for all impregnants except Halowax, which is measured at 65° C.
† Lowest temperature −40° C except for Halowax, which is −20° C.

It has been standard practice to rate d-c capacitors at a 40 deg cent ambient temperature on the basis that such a rating would provide sufficient factor of safety to withstand a life test of 1000 hours at twice the rated voltage at this temperature, which is equivalent to approximately 1 year of normal service conditions.

Design trends for electronic equipment toward smaller and smaller physical volume have resulted in ambient temperatures considerably in excess of the 40 deg cent value considered standard before World War II.

It has not been generally understood that increasing the ambient temperature above 40 deg cent required voltage derating for an equivalent life expectancy at the higher temperatures, and that the derating factors are a function of capacitor size. Capacitor size can best be evaluated for d-c ratings in terms of energy content in watt-seconds.

$$\text{Watt-second} = \tfrac{1}{2} CE^2$$

where C = capacitance in microfarads and E = d-c voltage in kilovolts.

This was a problem given joint industry and government appraisal during the war, and the findings resulted in a table of derating factors as a function of temperature and capaci-

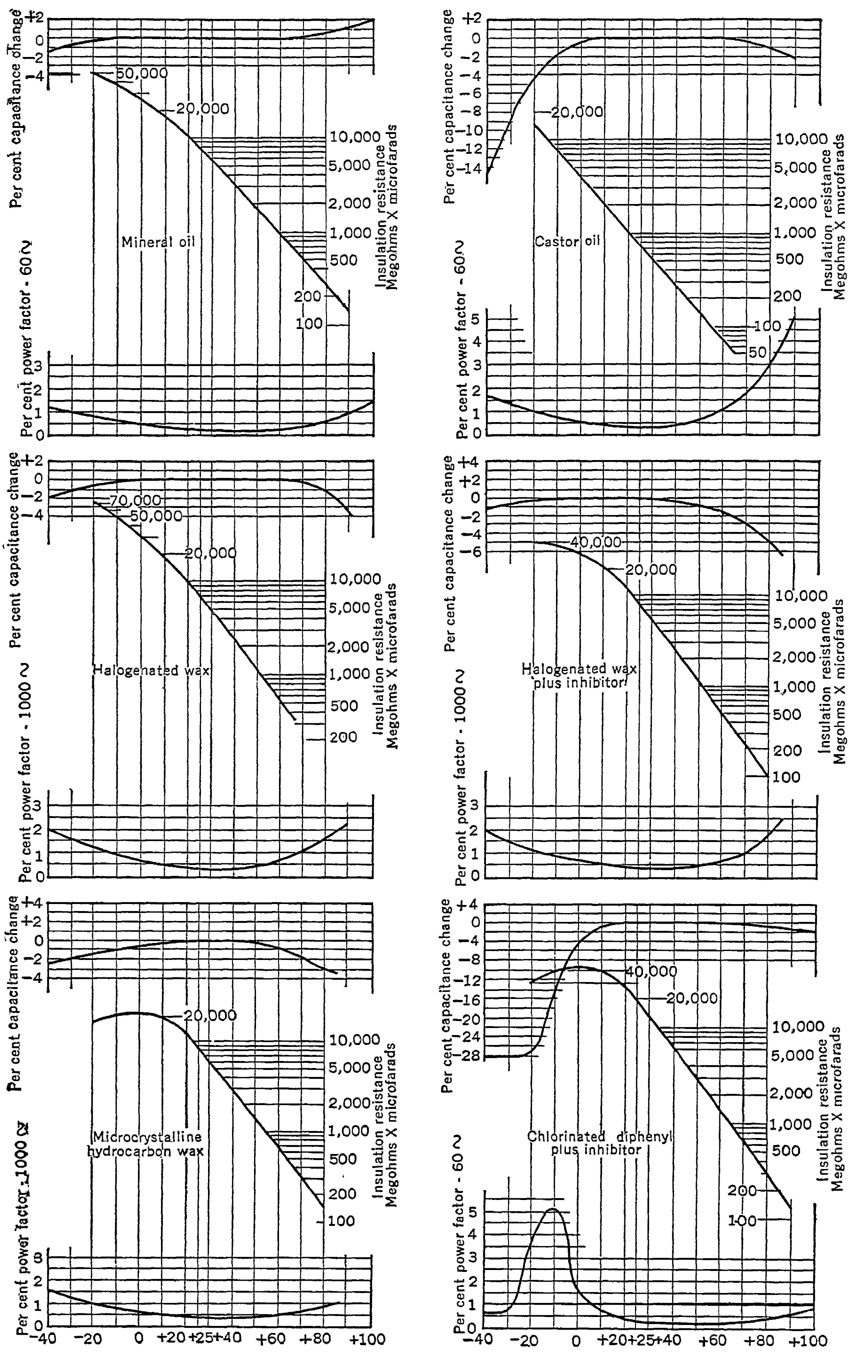


FIG. 8. Typical Electrical vs Temperature Characteristics of Paper-dielectric Capacitors with Various Impregnations (Courtesy of Solar Mfg. Corp.)

tor size. These data are shown in Fig. 9 and are extracted from data obtained in Proposed Joint Army-Navy Specification JAN-C-25 dated January 22, 1945, and published in *Communications* for August 1947.

These data show that a capacitor in the 0.5 watt-second class and rated for a given voltage at 40 deg cent must be voltage-derated to 95 per cent of its 40 deg cent rating at 85 deg cent or 60 per cent at 105 deg cent. Generally speaking, the life expectancy of a

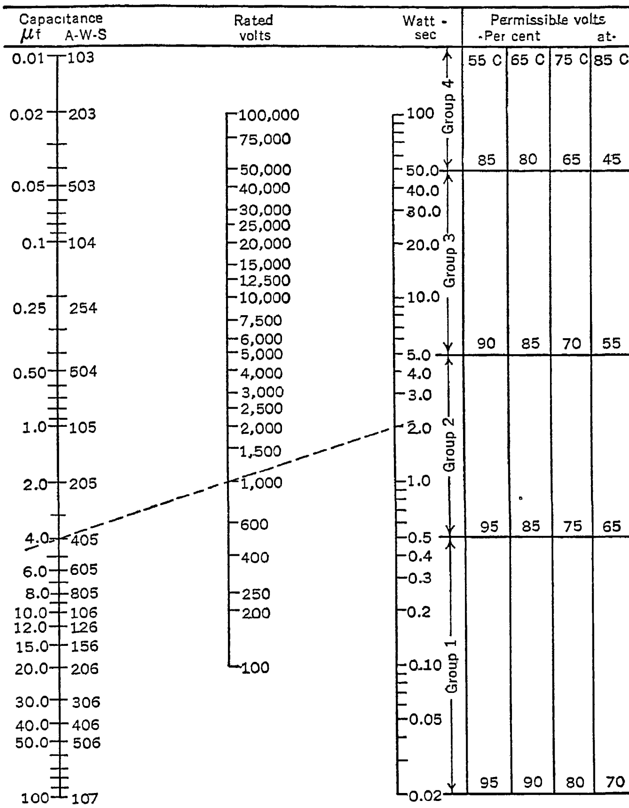


FIG. 9. Ratings of D-c Capacitors

capacitor without derating is halved for each 10 deg cent rise in temperature. For d-c voltage considerations, at a fixed temperature of 40 deg cent, the life expectancy is inversely proportional to the fifth power of the voltage.

D-c voltage ratings are not fixed values, but for a given insulation thickness they may be made variable depending upon the duty cycle and circuit conditions. A capacitor rated at 1000 volts d-c at 40 deg cent for continuous duty in a power-supply filter could be used in a photoflash circuit at a much higher voltage, such as 2000 volts d-c, and still have acceptable performance because of the lighter duty cycle. Life expectancy might increase to 100,000 flashes of the photoflash equipment.

This indicates that considerable flexibility may be used in rating capacitors provided that all the design requirements, operating conditions, and duty cycle are known. To illustrate, a 10-μf capacitor winding using three sheets of 0.0004 kraft tissue and impregnated in mineral oil would be normally rated at 1000 volts d-c at 40 deg cent continuous duty, at 1200 volts for an intermittent duty cycle of about 50 per cent, at 1500 volts intermittent duty for welder applications, and 2000 volts under a photoflash duty cycle. The d-c rating would have to be reduced to 600 volts for long-time (15 years) life expectancy.

Factors that affect capacitor ratings:

1. Microfarad value and tolerance.
2. Duty cycle, continuous or intermittent.

3. Ambient temperature range.
4. Ripple voltage, magnitude, and frequency.
5. Abnormal circuit voltages, such as no load voltage and peak charging voltage.
6. Discharge current, and nature of discharge, whether oscillatory, and, if so, whether critically damped.

It is well to remember that these factors that govern ratings are based on hermetically sealed capacitors which have been carefully dried, impregnated, and sealed. They do not apply to other constructions where life is limited by the vagaries introduced by moisture absorption.

METALLIZED PAPER CAPACITORS. The latest addition to the family of impregnated-paper capacitors is the MP type, in which the capacitor electrodes are deposited on the paper dielectric in very thin films, having a thickness range between 25 and 100 millimicrons. The thin film contributes the property of "self-healing" to capacitors, permitting the use of a single sheet of dielectric, which is not possible with conventional impregnated kraft paper designs. The combination of the extremely thin metallic film electrode and a single sheet of dielectric affords extremely compact designs for voltage ratings below 200 volts d-c or for 150 volts a-c.

Voltage ratings exceeding 200 volts employ a multiple-layer or interleaved paper dielectric of conventional construction. The volume saving is not as great as for the single-layer construction but is still considerable.

A new concept in capacitor rating is involved with MP capacitors, namely sparking voltage; this is defined as the lowest applied voltage that will cause continuous "self-healing" action to take place.

Table 2. D-c Voltage Ratings

D-c Working Voltage 25° C	1 Minute Flash Test 25° C	Sparking Voltage 25° C
200	300	400
400	600	900
600	900	1350

MP capacitors designed so that the maximum surge voltage encountered in service at the highest operating temperature does not exceed the sparking voltage are usable in all kinds of circuits without fear of their causing spurious noise. In the event of a transient voltage which would cause failure of a conventional capacitor type, there will be only a momentary arc discharge followed by the self-healing mechanism.

Insulation Resistance. The minimum insulation resistance of single-layer lacquered metallized paper capacitors will exceed 500 megohm microfarads or 2000 megohms at 25 deg cent. Interleaved unlacquered metallized paper capacitors will have a minimum insulation resistance of 1000 megohm microfarads or 6000 megohms at 25 deg cent, which compares favorably with conventional capacitor designs. The change in insulation resistance with temperature in metallized paper capacitors is similar to that of conventional mineral oil impregnated kraft paper designs, or there is approximately a 50 per cent decrease in insulation resistance for every 10 deg cent rise above 25 deg cent.

Commercial Specification References.

Joint Army-Navy Specification JAN-C-25, Capacitors, Direct Current, Paper Dielectric, Fixed (Hermetically Sealed in Metallic Cases).
Joint Army-Navy Specification JAN-C-91, Capacitors, Paper Dielectric, Fixed (Non-magnetic Cases).
RMA Standards Proposal 159.

20. MICA CAPACITORS

Mica capacitors are useful in electronic circuits because of their low a-c losses and their high electrical stability over a wide temperature range. These characteristics, along with the fact that they are constructed to very close capacitance tolerances, make them ideally suited for use in frequency-determining circuits.

The word "mica" is derived from the Latin "micare" meaning to sparkle. It is a group name for a number of aluminum silicate minerals which are characterized by the properties of high reflection and a basal cleavage so perfect that they may be split in laminae of the order of 0.0005 in. thick.

Of the eight distinct species of mica recognized by mineralogists, muscovite is the most important as far as mica capacitor manufacture is concerned. Muscovite is virtually unaffected by weathering, is not porous, is not decomposed by acids, and is negligibly affected by moisture. It will withstand relatively high voltage gradients. Voltage tests made with spherical electrodes show that films $^1/_{1000}$ in. thick frequently withstand 5000 volts d-c with no puncture. Muscovite, because of its extremely low power factor, in addition to its other desirable properties described previously, is an ideal dielectric for use in capacitor manufacture.

Both micas and foils or silvered electrode patterns must be precision outlined. Micas

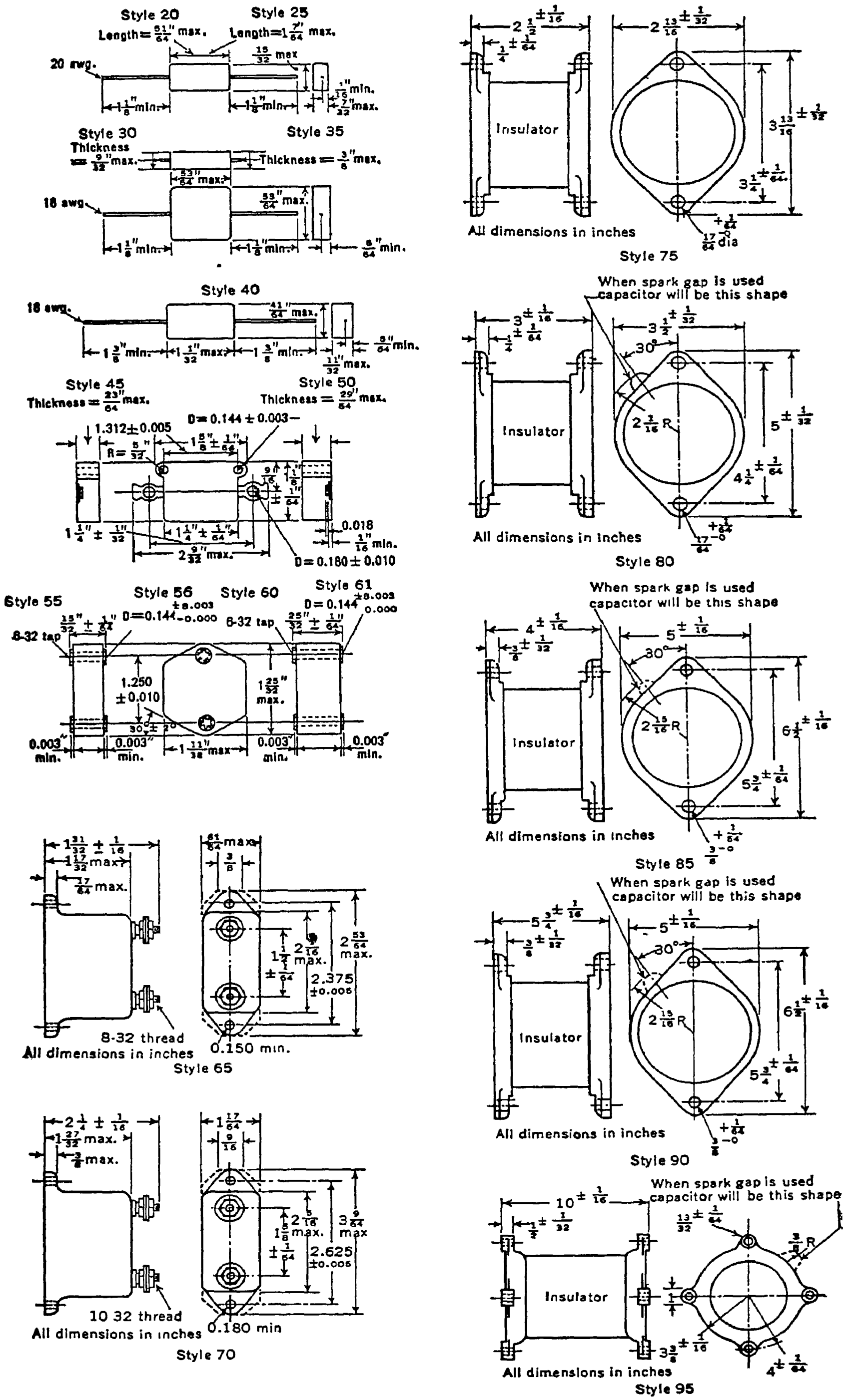


Fig. 10. Standard Outline Dimensions

are usually cut to within plus or minus 0.001 in. of specified lengths and widths and foils almost as accurately.

Sealing binders as well as metal clamps provide permanent positioning of the mica stack. This accuracy and permanency provide the stability of characteristics with respect to aging, frequency, and temperature which recommend mica capacitors for use in frequency-determining circuits or circuits that control reactance and phase and in precision measuring equipment.

Table 3. Capacitance Range for the Several Case Types

Case Type	From	To
20	5 μμf	1,000 μμf
25	5	1,500
30	470	10,000
35	3,300	10,000
40	100	10,000
45	17	10,000
50	2,000	27,000
55 and 56	22	30,000
60 and 61	100	47,000
65	47	100,000
70	47	100,000
75	47	100,000
80	47	100,000
85	47	100,000
90	100	100,000
95	100	10,000

Note: Capacitance values for 1000 μμf or less are measured at or referred to 500 kc/sec. Capacitance values greater than 1000 μμf are measured at or referred to 1000 kc/sec.

Standard commercial tolerance for molded-case types 20 to 61 is ±10 per cent and for potted types ±5 per cent; other tolerances are available as shown in Fig. 10.

Table 4. D-c Voltage or Peak Working Voltage Ratings for the Several Case Types

Style	Voltage Range
20, 25, 30, 35	300 and 500 volts
40	300, 500, 1000
45, 50, 55, 56, 60, 61	600, 1200, 2500
65	250, 500, 1000, 1500, 2000, 3000
70	500, 1000, 1500, 2000, 3000, 5000
75	1000, 1500, 2000, 3000, 4000, 6000
80	1500, 2000, 3000, 4000, 5000, 6000, 8000, 10,000
90	3000, 4000, 5000, 6000, 8000, 10,000, 12,000, 15,000, 20,000
95	15,000, 20,000, 25,000, 30,000, 35,000

Mica dielectric capacitors because of their low losses and high Q are useful at high frequencies and for high current circuits.

The mechanical forms of mica dielectric capacitors vary from small phenolic molded cases for receiver applications to large ceramic insulated housings required by high-current, high-voltage-transmitting circuits.

The mica capacitor is the only one of the many capacitor types that has obtained general industry standardization for case styles and electrical ratings.

Figure 10(*a*) shows outline dimensions for molded-case types, Styles 20, 25, 30, 35, 40, 45, 50, 55, 56, 60, and 61; Figs. 10(*b*) and (*c*) show outline dimensions for the potted-case types, Styles 65, 70, 75, 80, 85, 90, and 95.

Table 5. Classification

Designation	Temperature Coefficient Not more than	Capacitance Drift Not more than
Class A..........	±1000 ppm	±(5% + 1 μμf)
Class B..........	± 500 ppm	±(3% + 1 μμf)
Class C..........	± 200 ppm	±(0.5% + 0.5 μμf)
Class I..........	{+ 150 / − 50 ppm}	±(0.3% + 0.2 μμf)
Class D..........	± 100 ppm	±(0.3% + 0.1 μμf)
Class J..........	{+ 100 / − 50 ppm}	±(0.2% + 0.2 μμf)
Class E..........	{+ 100 / − 20 ppm}	±(0.1% + 0.1 μμf)
Class G..........	{0 / − 50 ppm}	±(0.1% + 0.1 μμf)

Note: Characters D, J, E, and G require individual tests of each capacitor and should be considered for use only where extreme stability and accuracy are required.

Capacitance. Capacitance values for mica capacitors are expressed in micro-microfarads.

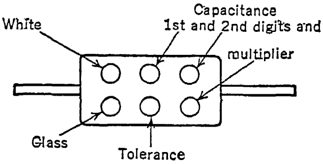


FIG. 11. RMA Capacitor Color Code (Proposed)

Color Marking of Molded Types. A convenient system of six-color marking is being used to identify molded mica case types. Two RMA color codes are in use, one as a standard and the other proposed, but industry standardization is expected to make the old system, which does not provide for identification of the class designation, obsolete.

Figure 11 shows the arrangement of the two rows of colors, and the significance of each color is shown in Table 6.

Radio-frequency Current Ratings. The potted-case type of mica capacitor is intended primarily for use in frequency-determining circuits or those requiring the capacitor to handle appreciable amounts of r-f current. Typical curves illustrating the current-carrying capacity for a frequency range of 0.1 to 30 megacycles of mica dielectric capacitors housed in potted ceramic cases are shown by Fig. 12.

Table 6

Color	Numerical Significance	Decimal Multiplier	Capacitance Tolerance	Class Designation
Black.....	0	1	20%	A
Brown....	1	10		B
Red.......	2	100	2%	C
Orange....	3	1,000	3%	D
Yellow....	4	10,000		E
Green.....	5		5%	
Blue......	6			
Violet.....	7			
Gray......	8			I
White.....	9			J
Gold......		0.1		
Silver.....		0.01	10%	

It is interesting to note the increasing current loading with frequency for the 0.0001-$\mu\mu f$ capacitor in comparison with the decreasing current rating with increasing frequency for capacitance values greater than 0.001 $\mu\mu f$, which is due to the inductance introduced in the series-paralleling of the mica sections which constitute the capacitor stack for the larger capacitance values.

The current-carrying capacity of mica dielectric capacitors can be materially increased by immersing the foil-mica stack in silicon oil and removing the heat generated by means of cooling coils.

Manufacturers' published ratings are based on an ambient temperature of 40 deg cent, and the current ratings must be derated for higher ambient temperatures as shown in Table 7.

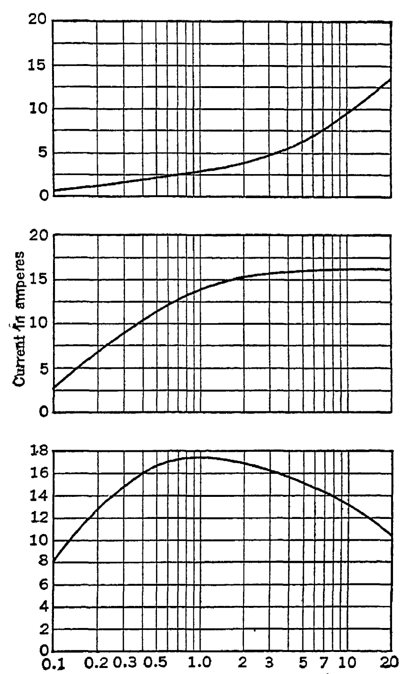


FIG. 12. Current-carrying Capacity for 10 Deg Cent Temperature Rise (Courtesy of Solar Mfg. Corp.)

Table 7. Current Deratings with Temperature

Characteristic	Temperature Range	Current Derating Factor
B, C	41 to 50 deg cent	0.95
B, C	51 to 60	0.85
B, C	61 to 70	0.70
D, E, G	40 to 70	0.50

Commercial Specification References.

Joint Army-Navy Specification JAN-C-5, Capacitors, Mica-Dielectric, Fixed.
RMA Standards Proposal 158A.

21. CERAMIC DIELECTRIC CAPACITORS

Ceramic capacitors are neither new nor the result of any accidental discovery. They are the direct result of long research in the early 1900's when German scientists noted the unusual characteristics of titanate ceramic materials.

Europe has never had a domestic source of mica and was, therefore, faced with a serious shortage before World War I. This shortage focused attention and accelerated research into the possibilities of developing a suitable substitute. This problem in ceramic research took German scientists many years to solve. In the early 1930's, a ceramic titanate material was finally developed that was controllable in production quantities and, at the same time, would retain stable characteristics. These ceramic dielectric units were quickly used in substantial quantities by European radio and electronic industries.

The ceramic dielectric because of its negative temperature coefficient was used in temperature-compensating capacitors in oscillator circuits. As more experience was gained and the basic characteristics became more clearly understood, other forms of ceramic capacitors were used, including higher capacitances, such as by-pass and coupling types for transmitting and other high-voltage and high-current capacitors.

The electronics industry in the United States was slow to recognize the possibilities of ceramic capacitors and to utilize the general existing knowledge of European ceramic dielectric materials because of the abundant supplies of mica which were readily available. A second retarding factor was the general feeling that ceramic dielectrics in the titanate group would be far too costly for our mass-production methods.

Centralab, a division of Globe Union, Inc., in the early 1930's, initiated a research program to investigate the availability of raw materials and the possibility of producing ceramic capacitors similar to the European types. Abundant domestic supplies of raw materials were located which exhibited characteristics superior to those of the European materials.

The first group of capacitors was offered in capacitance value up to 1000 $\mu\mu$f and in controlled temperature coefficient varying from a positive change of 100 ppm to a negative change of 750 ppm. The dielectric constant of these materials varied from 50 in the zero temperature coefficient group to a maximum of 95 in the group having a negative temperature coefficient of 750 ppm.

The use of ceramic capacitors was greatly accelerated during World War II because of the increased demand for substitutes for mica capacitors, which were in short supply owing to a shortage of high-quality mica.

The fact that ceramic capacitors possess low losses at ultra-high frequencies makes them ideally suited for application where other types of dielectrics are not satisfactory. They are available in both tubular and disk constructions. The disk constructions employ a "feed-through" terminal arrangement which reduces lead inductance to the absolute minimum.

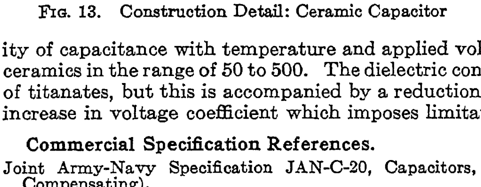


FIG. 13. Construction Detail: Ceramic Capacitor

A typical construction for the tubular ceramic capacitors is shown in Fig. 13. In this construction, the capacitance is controlled by the dielectric constant of the ceramic, by the length, diameter, and wall thickness of the ceramic tube, and by the silvered area of the electrodes. Stability of capacitance with temperature and applied voltage is obtained with the use of low-K ceramics in the range of 50 to 500. The dielectric constant may be increased through the use of titanates, but this is accompanied by a reduction in capacitance stability and a marked increase in voltage coefficient which imposes limitations on the use of the high-K bodies.

Commercial Specification References.

Joint Army-Navy Specification JAN-C-20, Capacitors, Ceramic-Dielectric, Fixed (Temperature-Compensating).
RMA Standards Proposal 157, Ceramic Dielectric Capacitors.

22. ELECTROLYTIC CAPACITORS

Electrolytic capacitors employ solid dielectric media on which an oxide film, produced electrochemically, is the dielectric in the presence of a d-c polarizing voltage and an ionic conducting medium.

Investigations have shown that the oxides of tantalum and aluminum exhibit desirable characteristics. Tantalum possesses an oxide with high dielectric constant and low leakage, but economic factors have limited its use. Its principal limitations are voltage rating and mechanical construction resulting from the use of sulfuric acid as the electrolyte. Aluminum is the metal used in all other electrolytic capacitors because of its low price and excellent film-forming characteristics.

An oxide film can be formed on aluminum by electrolytic means by immersing a ribbon of aluminum foil in an aqueous solution of boric acid and sodium borate and passing an electric current through the solution with the aluminum forming the positive pole or anode. Electrolysis of the solution causes oxygen to be generated at the positive pole, oxidizing the surface of the aluminum. The film thickness is a function of the d-c formation voltage.

The extremely thin oxide film formed on the aluminum anode offers a very high resistance to further passage of current if the applied voltage is not increased above the film formation voltage. A cell of this nature inserted in a container containing an aqueous electrolyte takes the form of the so-called wet electrolytic capacitor. The aluminum oxide film acts as the dielectric, the electrolyte as the cathode, and the container as the contact medium for the cathode.

The electrolytic capacitor has a very high capacitance per unit volume as compared to other types of capacitors, such as the impregnated-paper or mica dielectric types.

The primary reason why an electrolytic capacitor gives a high capacitance per unit volume is the extreme thinness of the dielectric or oxide film. The thickness of the oxide film covering the aluminum electrode is approximately 2×10^{-5} in., and the dielectric constant K of the oxide layer produced is high (approximately 10) as compared with 2.5 for a mineral oil impregnated paper capacitor.

There are two types of electrolytic capacitors, depending upon the physical characteristics of the electrolytes: the *wet type*, which uses an aqueous electrolyte; and the *dry type*, which uses a viscous or paste electrolyte.

ANODE FOIL TREATMENTS. The anode foil employed in electrolytic capacitors may assume several forms, depending upon design considerations:

1. *Plain foil*, where the oxide film is electrochemically formed on the surface of the aluminum foil without any previous treatment of the foil surface.
2. *Etched foil*, where the surface of the aluminum foil is first treated chemically or electrochemically to erode the surface, thereby increasing the superficial area prior to the film-forming procedures.
3. *Sprayed gauze*, where an inert carrier such as chemically pure cotton gauze is mechanically coated with aluminum by means of metal spraying.

The etched foil electrode is the most common one in both wet and dry types of electrolytic capacitors, although the sprayed gauze construction is becoming more common in the dry electrolytic capacitor types.

The hydrochloric acid etched anode foil construction is used in place of plain foil because it gives effective surface areas much greater than those obtained from plain foil and thereby cuts the physical size.

CATHODE OR COLLECTOR FOIL TREATMENTS. Filming of the cathode foil is a very interesting phenomenon which has been observed in very high gain foils in circuits of high ripple currents and which must be carefully considered in the use of electrolytic capacitors. The mechanism that causes cathode film formation can be explained by the fact that the cathode foil under high ripple current conditions is subject to a reversal of ripple current due to the charging of the capacitor on the conducting part of the cycle and the discharging of the capacitor into the load on the non-conducting part of the cycle even though the cathode never becomes positive with respect to the anode polarizing potential, as illustrated by Fig. 14. As can be seen from the diagram, on one part of the cycle the capacitor is charging and during the other half-cycle the capacitor is discharging through the load. The result is that the cathode becomes anodized to the value of the ripple voltage impressed across the capacitor. The value of this potential is a function of the power-supply regulation, the capacitor impedance, and the magnitude of the load current. When the capacitance of the cathode is reduced by formation, it produces in effect a low instead of a high capacitance in series with the anode capacitance and reduces the capacitance of the capacitor, depending upon the degree of cathode film formation.

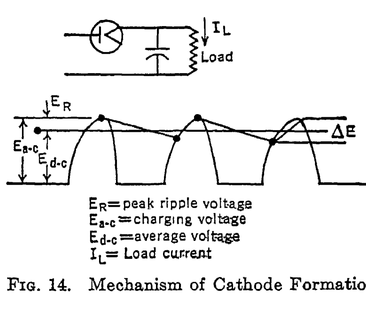


FIG. 14. Mechanism of Cathode Formation

Etched Cathodes. Most commercial designs employ etched cathode foil for low-voltage sections (less than 25 working volts d-c) and for other ratings where ripple current is high. Etched cathode foil is used for the following reasons:

1. An etched cathode foil increases the effective cathode surface area, which reduces the ripple current density to such a value that the cathode film does not build up to a value exceeding the initial thickness of the cathode film.
2. The increased cathode surface will afford a much higher cathode capacitance, and, the higher the value of the cathode capacitance, the smaller will be the reduction in initial

anode capacitance because of the series connection of the cathode and anode capacitance. For example, a 100-μf 15 working volt d-c electrolytic capacitor with a plain foil cathode has a cathode capacitance of approximately 120 μf, which would cut the total capacity of the unit to 54.5 μf. However, if an etched cathode foil was used, the cathode capacitance would be approximately 1200 μf, and the resultant total capacitance of the capacitor would be 93 μf.

Table 8 gives the maximum root mean square ripple currents recommended in RMA Standards Proposal 160A for various capacitance and voltage ratings.

Table 8. Maximum rms Ripple Current in Milliamperes at 120 Cycles

Micro-farads	15 v	25 v	50 v	150 v	250 v	300 v	350 v	400 v	450 v
10					80	90	145	150	150
20				90	160	180	180	180	180
30				135	180	200	200	200	200
40				190	190	200	200	200	200
50				200	200	200	200	200	200
60			135	200	200	200	200	200	200
70			150	200	200	200	200	200	200
80			200	200	200	200	200	200	200
90			200	200	200	200	200	200	
100		200	200	200	200	200	200		
200		300	300	300	300	300			
300		400	400	400					
400		500	500						
500	500	600	625						
600	550	650	700						
700	580	700	750						
800	625	750	900						
900	700	850							
1000	700	850							
1500	850	1000							
2000	1000								

It is very important that electrolytic capacitors should not be subjected to ripple currents in excess of the values listed in Table 8; otherwise, the capacitors will overheat and the life will be appreciably reduced.

CHARACTERISTICS OF ELECTROLYTIC CAPACITORS. A knowledge of the basic characteristics of electrolytic capacitors is essential for their correct use. These are:

Table 9. Commercial Capacitance Tolerances

D-c Working Voltage	Capacitance Tolerance
0– 50	−10; +250%
51–350	−10; +100%
351–450	−10; +50%

Capacitance. The capacitance of a dry electrolytic capacitor is determined by the surface area of the anode and the dielectric thickness by the formation voltage. With etched foil construction, the increase in superficial area or gain varies inversely with the formation voltage. There is a problem of capacitance control with formation voltage, since for lower voltages capacitance per unit area varies considerably. See Table 9.

Leakage Current. The leakage-current characteristic of a dry electrolytic capacitor represents the amount of direct current flowing through the capacitor with its rated polarizing voltage applied, and does not include the momentary charging current. This leakage current is an indication of the quality of the anode film. Commercial specifications for permissible leakage current may be determined by the formula:

$$I = KC + 0.3$$

where I is the d-c leakage in milliamperes, K is the constant as shown in Table 10, and C is the rated capacitance in microfarads. The leakage current is determined after application of rated d-c working voltage.

Table 10. Leakage-current Constant

D-c Rated Voltage	K
3 to 100	0.01
101 to 250	0.02
251 to 350	0.025
351 to 450	0.04

D-c Working Voltage. The d-c working voltage is the maximum d-c voltage the capacitor will stand under continuous operation within its normal temperature range.

Peak Working Voltage. The peak working voltage represents the d-c voltage plus the peak a-c ripple voltage; it refers to a continuous operating condition and should not be confused with surge voltage.

Surge Voltage. The surge voltage is a short-time d-c voltage rating that exceeds the peak working voltage and approaches the film formation voltage. This voltage is limited by the internal heating of the capacitor caused by the rapid increase in leakage current as shown by Fig. 17. Usage has established surge-voltage ratings for various d-c working voltages which are listed in Table 11 for a 1000-ohm circuit regulation resistance.

Power Factor. For all practical purposes, the power factor of an electrolytic capacitor is the ratio between equivalent series resistance and the capacitance reactance at a given frequency. It is expressed in percentage and indicates the energy consumed by the capacitor.

Equivalent Series Resistance. Equivalent series resistance is a more useful characteristic in mathematical equations relating to electrolytic capacitors. The equivalent series resistance represents the total losses. r_s = watts/i^2, where r_s = equivalent series resistance and i = leakage current. The total losses in an electrolytic capacitor consist of: (1) dielectric loss of oxide film, (2) electrolyte resistance, and (3) contact resistance. The combined effect of these losses is expressed as the equivalent series resistance value necessary to produce an i^2r loss of the same magnitude.

Table 11. Single-voltage Ratings.

D-c Rated Voltage	D-c Surge Voltage
3	4
10	12
15	20
25	40
50	75
150	185
250	300
300	350
350	400
400	450
450	500

A convenient figure of merit for evaluating losses in electrolytic capacitors is the P factor, which is expressed as the product of the rated capacitance in microfarads and equivalent series resistance in ohms, as measured on a polarized capacitance bridge at a frequency of 120 cycles per second. Commercial capacitors have P factors which do not exceed the values shown in Table 12 for the several standard d-c voltage ratings. When the high- and low-voltage sections are combined into a single capacitor winding, the electrolyte for the high-voltage section determines the P factor for the low-voltage section and raises the P factor over what would be obtained with a low-voltage electrolyte. This makes it necessary to double the P factor (Table 12) for the low-voltage section in combination with sections having d-c ratings exceeding 150 volts.

Table 12. *P* **Factors**

Rated Voltage	P
3	3000
10	1500
15	1200
25	500
50	400
150	300
250	250
300	250
350	250
400	250
450	250

Temperature Effects. Capacitance, series resistance, power factor, and impedance of electrolytic capacitors are all somewhat affected by temperature. Figure 15 shows capacitance and power factor vs. temperature for typical commercial electrolytic capacitors when operated over a temperature range of −20 to 85 deg cent.

Aircraft and other special industrial applications require dry electrolytic capacitors which will maintain constancy of characteristics for temperatures as low as −40 deg cent. Special electrolytes are available which meet these requirements; see Fig. 16.

Dry electrolytic capacitors are not recommended for use in ambient temperatures exceeding 85 deg cent because of rapid drying out of the electrolyte.

Leakage current increases with temperature, as shown by curve Fig. 17 for a high-voltage filter capacitor and low-voltage by-pass section combined in a single winding.

Dry electrolytic capacitors intended for operation at ambient temperatures of 85 deg cent require the electrolyte to be heat-stabilized to prevent the leakage current from increasing to the point where the capacitor overheats and fails. The dotted curve of Fig. 18 shows the reduction in leakage current at rated d-c voltage for various temperatures in comparison with the leakage current for these same temperatures before heat stabilizing.

Early failure of dry electrolytic capacitors in electronic equipment is often due to failure to allow for excessive temperature in the electrolytic capacitor specifications.

R-f Impedance. Multiple-section concentrically wound dry electrolytic capacitors employ a common cathode which gives rise to coupling in the cathode circuit because of the common current paths as shown by Fig. 19. In circuits where common coupling causes circuit unstability, a swedged cathode is employed. This construction effectively cuts the common r-f impedance to a negligible value by the extension of the cathode foil with the turns swedged together, which, in effect, gives a non-inductive cathode construction since each cathode or ground terminal is directly under its corresponding anode terminal. This construction is similar as far as the cathode is concerned to the non-inductive winding shown in Fig. 6 for impregnated-paper capacitors.

Gas Pressure. Electrolytic capacitors should be supplied with a liberal space for gas expansion. This extra space is to accommodate any sudden generation of gas which may

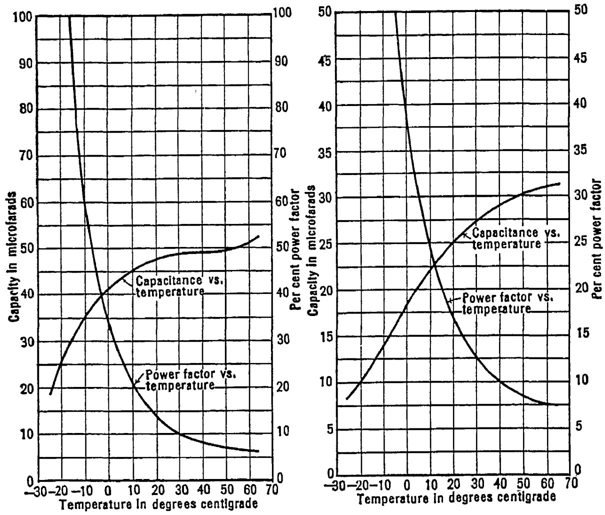


FIG. 15. Capacitance and % Power Factor vs Temperature of Electrolytic Capacitor (Courtesy Solar Mfg. Corp.)

be liberated as the result of improper uses of the capacitor. In addition, most electrolytics are supplied with built-in vents which prevent capacitors from exploding when they are improperly used, as being accidentally connected across alternating current.

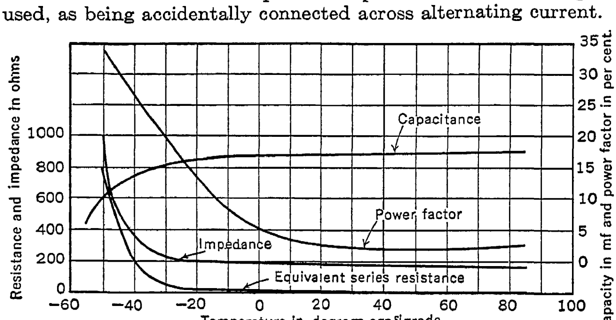


FIG. 16. Variation of Characteristics of Low-temperature Electrolytic Capacitor with Temperature (16 μf—350 WVDC—Courtesy of Solar Mfg. Corp.)

Polarity. *Polarized* types of dry electrolytic capacitors are designed for use in d-c or intermittent d-c circuits produced by rectifying alternating current. D-c polarized capacitors should not be subjected to reversed polarity, as the heavy current passing through the capacitor under this condition will raise its internal temperature and seriously damage it.

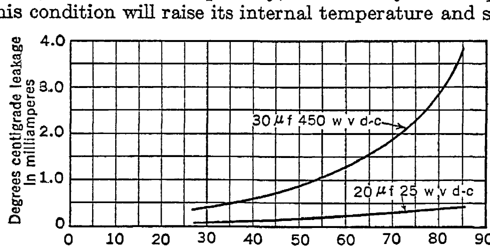


FIG. 17. D-c Leakage vs Temperature—Dry Electrolytic Capacitors

Non-polarized Types. This term applies to dry electrolytic capacitors constructed with two formed electrodes so that they function equally well with direct current impressed on either electrode irrespective of polarity. This does not mean that they can be used on alternating current continuously. Non-polarized capacitors are for applications where the d-c voltage supply might become reversed and remain so indefinitely. A non-polarized capacitor is equivalent to two polarized capacitors connected in series opposition.

A-c Motor-starting Capacitors. Non-polarized dry electrolytic capacitors may be used for intermittent duty in a-c circuits such as for motor starting, and they are known as motor-starting capacitors. They are ideal for intermittent duty when used within the manufacturers' limits for operating voltage, temperature, and duty cycle.

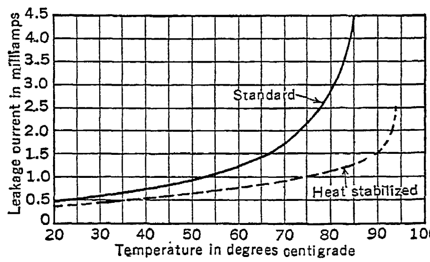


FIG. 18. Effect of Heat Stabilization of Electrolyte (30 μf—450 WVDC—Courtesy of Solar Mfg. Corp.)

Since the duty cycle is based on internal heating of the capacitor, it is possible to vary the number of application periods of voltage with the period of duration of voltage so that the product is a constant. The manufacturers' guarantee is usually twenty 0.5-sec starts per hour.

The normal maximum operating temperature of these capacitors is 65 deg cent. They may be successfully operated up to 85 deg cent provided that the duty cycle and maximum voltage conditions are adjusted accordingly. However, it must be noted that operating at higher than normal temperatures decreases the life considerably as it dries out the electrolyte at an accelerated rate. Operation of capacitors at low temperatures cannot harm capacitors because any change in characteristics is only temporary at subzero temperatures. The increase in power factor represents an increased resistance loss when the capacitor is in operation, and this creates sufficient heat to warm up the capacitor and quickly return it to normal operating conditions.

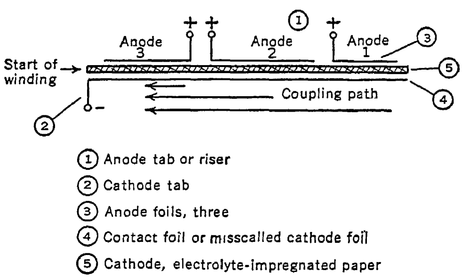


FIG. 19. Schematic Diagram of Three Section Dry Electrolytic Capacitor Winding

Commercial Specification Reference. Where more specific information and data are required for testing methods and procedures and for specific capacitor ratings, the following are suggested:

Joint Army-Navy Specification JAN-C-62, Capacitors, Dry Electrolytic, Polarized.
RMA Standards Proposal 160, Polarized Dry Electrolytic Capacitors.

BIBLIOGRAPHY

A.I.E.E. Standards for Capacitors, Approved Standard 18, June 1934.
Radio Instruments and Measurements, *Circular of the Bureau of Standards* 74, Government Printing Office, 1924.
Maloff, I. G., Mica Condensers in High-frequency Circuits, *Proc. I.R.E.*, Vol. 20, 647 (April 1932).
Kouenhoven, W. B., and Lemmon, C. L., Phase Defect Angle of an Air Capacitor, *J. A.I.E.E.*, Vol. 49, No. 11, 945 (November 1930).
Field, R. F., An Equal-arm Capacitance Bridge, *General Radio Experimenter*, Vol. 4, No. 8, 1 (January 1930).
Burke, C. T., Substitution Method for the Determination of Resistance of Inductors and Capacitors at Radio Frequencies, *Trans. A.I.E.E.*, Vol. 46, 483 (May 1927).
Gemant, A., *Liquid Dielectrics*, John Wiley, 1933.
Morgan, S. O., and White, A. H., The Dielectric Constant and Power Factor of Rosin Oil and Ethyl Abietate, *J. Franklin Inst.*, Vol. 213, No. 3, 313 (March 1932).
Hoch, E., Power Losses in Insulating Materials, *Bell Sys. Tech. J.*, Vol. 1, No. 2, 110 (November 1922).
Benedict, R. R., Behavior of Dielectrics, *Trans. A.I.E.E.*, Vol. 49, 739 (April 1930).
Lewis, A. B., Hall, E. L., and Caldwell, F. R., Some Electrical Properties of Foreign and Domestic Micas, *Bur. Stand. J. Research*, Vol. 7, No. 2, 403 (August 1931).
Siegmund, H. O., The Aluminum Electrolytic Condenser, *Bell Laboratories Reprint* 349, *Trans. Electrochem. Soc.*, Vol. 53, 203 (1928).
Godsey, F. W., Jr., Film Characteristics of Electrolytic Condensers, *Trans. A.I.E.E.*, Vol. 51, 432 (June 1932).

Godsey, F. W., Jr., A-C Capacity of Electrolytic Condensers, *Trans. Electrochem. Soc.*, Vol. 61, 515 (April 1932).

Godsey, F. W., Jr., Cathodic Films in Electrolytic Condensers, *Trans. Electrochem. Soc.*, Vol. 63, 223 (1933).

Godsey, F. W., Jr., Potential Gradients in Anodic Films, *Trans. Electrochem. Soc.*, Vol. 61, 549 (1932).

Lilienfeld, Appleton, Smith, and Nieh, Studies of Fully Organized Anodic Layers on Aluminum, *Trans. Electrochem. Soc.*, Vol. 61, 531 (1932).

Godsey, F. W., Jr., Power Losses in Electrolytic Condensers, *Trans. A.I.E.E.*, Vol. 51, 439 (June 1932).

Christopher, A. J., and Kater, J. A., Mica Capacitors for Carrier Telephone Systems, *Trans. A.I.E.E.*, Vol. 65, 670 (October 1946).

Murphy, E. J., and Morgan, S. O., The Dielectric Properties of Insulating Materials, *Bell Sys. Tech. J.*, Vol. 16, 493 (October 1937).

MacLeod, H. J., The Variation with Frequency of the Power Loss in Dielectrics, *Phys. Rev.* (2), Vol. 21 (January 1923).

Brotherton, M., Paper Capacitors under Direct Voltage, *Proc. I.R.E.*, Vol. 32, 139 (March 1944).

McLean, D. A., Edgerton, L., Kohman, G. T., and Brotherton, M., Paper Dielectrics Containing Chlorinated Impregnants, *Industrial and Eng. Chem.*, Vol. 34, 101 (January 1942).

Palmer, H. B., Capacitance of a Parallel Plate Capacitor by the Schwartz-Christoffel Transformation, *Trans. A.I.E.E.*, Vol. 56, 363 (March 1937).

Reed, M., Effect of Stray Capacities to Ground in Substitution Measurements, diagrams, *Wireless Eng.*, Vol. 13, 248 (May 1936).

Field, R. F., and Sinclair, D. B., Method for Determining the Residual Inductance and Resistance of a Variable Air Condenser at Radio Frequencies, bibliography, diagrams, *Proc. I.R.E.*, Vol. 21, 255 (February 1936).

McDonald, L. J., Contours of Capacitor Rotor Plates, *Electronics*, Vol. 18, 126 (March 1945).

Boella, M., Direct Measurement of the Loss Conductance of Condensers at High Frequencies, *Proc. I.R.E.*, Vol. 20, 421 (April 1938).

Direct-reading Condenser for Substitution Measurements, *Gen. Radio Exp.*, Vol. 10 (March 1936).

Michaelson, H. B., Gas-filled and Vacuum Capacitors, illustrations, *Electronics*, Vol. 17, 124 (September 1944).

Sinclair, D. B., High-frequency Model of Precision Condenser, *Gen. Radio Exp.*, Vol. 12 (October–November 1939).

Griffiths, W. H. F., Law Linearity of Semicircular Plate Variable Condensers, diagrams, *Wireless Eng.*, Vol. 22, 107 (March 1945).

Field, R. F., and Sinclair, D. B., Method for Determining the Residual Inductance and Resistance of a Variable Air Condenser at Radio Frequencies, bibliography, diagrams, *Proc. I.R.E.*, Vol. 24, 255 (February 1936).

Green, A. P., and McComb, C. T., Resonance in Mica Capacitors, *Electronics*, Vol. 17, 119 (March 1944).

Brinkmann, C., Self-discharge and Time Constant of the High-voltage Oiled-paper Condenser, abstract, *Wireless Eng.*, Vol. 20, 449 (September 1943).

Schick, W., Temperature Coefficient of Capacitance; Its Measurement in Small Radio Condensers, bibliography, illustrations, diagrams, *Wireless Eng.*, Vol. 21, 65 (April 1944); Discussion, T. J. Rehfisch, Vol. 21, 175 (February 1944).

Coursey, P. R., Thermal Stability of Condensers; Ceramic Dielectrics and Their Use at Low Temperatures, *Wireless Eng.*, Vol. 15, 247 (May 1938).

Floyd, G. H., Vacuum Capacitors, illustrations, diagrams, *Proc. I.R.E.*, Vol. 32, 463 (August 1944).

Proctor, R. F., Variable Air Condensers; Determination of Their Residual Parameters, diagrams, *Wireless Eng.*, Vol. 17, 257 (June 1940).

Attwood, S. S., and Bixby, W. H., Breakdown and Time-Lag of Dielectric Materials, *J. Franklin Inst.*, March 1943.

Balsbaugh, J. C., Assaf, A. G., and Oncley, J. L., Dielectric Properties of Hydrocarbons and Hydrocarbon Oils, *Industrial and Eng. Chem.*, January 1942, pp. 92–100.

Barringer, L. E., The Relation of Chemical and Physical Structure to Dielectric Behavior, *Trans. Electrochem. Soc.*, Vol. LXV (1934).

Burnett, J. H., Liquid, Viscous, and Solid Dielectrics, *Electric Manufacturing*, August 1943.

Clark, F. M., The Development and Application of Synthetic Liquid Dielectrics, *Trans. Electrochem. Soc.*, Vol. LXV (1934).

Clark, F. M., and Raab, E. L., Electrical Stability of Mineral Oil-Treated Dielectrics, *Industrial and Eng. Chem.*, January 1942, pp. 110–116.

McLean, D. A., and Egerton, L., Paper Capacitors Containing Chlorinated Impregnants. Stabilization by Anthraquinine, *Industrial and Eng. Chem.*, Vol. 37 (January 1945).

Cornell, J. I., Metallized Paper Capacitors, *The Solar System*, Vol. IV, No. 4 (November-December 1946), Solar Mfg. Corp.

Berberich, L. J., Fields, C. V., Marbury, R. E., Characteristics of Chlorinated Impregnants in D-c Paper Capacitors, *A.I.E.E. Technical Paper* 44–165 (May 1944).

Karplus, Edward, Wide Range Tuned Circuits and Oscillators for High Frequencies, *Proc. I.R.E.*, Vol. 33, 426 (July 1945).

Piper, J. D., Kerstein, N. A., and Fleiger, A. G., Oil Impregnated Paper, *Industrial and Eng. Chem.*, September 1937, pp. 1040–1043.

Siegmund, H. O., The Aluminum Electrolytic Condenser, *Bell Sys. Tech. J.*, January 1929.

White, A. H., and Morgan, S. O., The Dielectric Properties of Chlorinated Diphenyls, *J. Franklin Inst.*, November 1933, pp. 635–644.

Whitehead, J. B., Liquid Insulators, *Trans. Electrochem. Soc.*, Vol. LXV (1934).

Zimmerman, C. I., The Aluminum Electrolytic Condenser, *Trans. Am. Electrochem. Soc.*, Vol. V (1904).

Berberich, L. J., and Friedman, Raymond, Stabilization of Chlorinated Diphenyl in Paper Capacitors, *Industrial and Eng. Chem.*, Vol. 40, 117 (January 1948).

Golding, E. W., *Electrical Measurements and Measuring Instruments*, Sir Isaac Pitman & Sons, Ltd., London.

Georgiev, Alexander M., *The Electrolytic Capacitor*, Murray Hill Books.

Brotherton, M., *Capacitors—Their Use in Electronic Circuits*, D. Van Nostrand Co.

Nersey, Philip R., *Electrolytic Condensers*, Chapman Hall, 2nd Ed., London, 1939.

Deeley, Paul McKnight, *Electrolytic Capacitors*, Cornell-Dubilier, 1938.

Coursey, P. R., *Electrical Condensers, Their Construction, Design and Industrial Uses*, Pitman, 1927.

Schwaiger, A., *Theory of Dielectrics*, John Wiley, 1932.

Whitehead, J. B., *Impregnated Paper Insulation*, John Wiley and Sons, 1935.

SECTION 4

ELECTRON TUBES

ELECTRON TUBES

THERMIONIC VACUUM TUBES

By A. P. Kauzmann

As dealt with in articles 1–6, vacuum tubes in operation are characterized by a source of electron emission; the conduction through a vacuous space, which may or may not contain sufficient gas to affect the conduction, of a current between the source of emission and one or more other electrodes by means of the emitted electrons; and the varying of this current by means of variations in electrode potential to produce an electrical response in an associated circuit. Thus, phototubes and cathode-ray tubes, though strictly vacuum tubes, are excluded (see Section 15). Grid-controlled gas-discharge tubes (see articles 13–21) and mercury-pool-type rectifiers (see article 21) are treated later in the section because their special properties and applications require special treatment.

1. PRINCIPLES OF OPERATION

HIGH-VACUUM TUBES. The simplest form of *high-vacuum tube* is the *diode*, consisting of a *thermionic cathode* and an *anode.* The cathode is heated to a temperature, depending on its nature, at which electrons are emitted from its surface into the surrounding vacuous space. When the anode is placed at a potential positive with respect to the cathode some of these emitted electrons are caused to flow to the anode under the influence of the electrostatic field, thus constituting a current flowing from cathode to anode in the external circuit, since the electrons are negatively charged. Those electrons which are not drawn to the anode return to the cathode.

Because of the mutual repulsion between the like charges of the electrons, only a definite number of them may be accommodated in the space between cathode and anode at any given anode potential, and therefore the current flowing is definitely limited. This limiting effect is called *space charge*, and the current is said to be *space-charge limited.* As the anode potential is increased the current increases until the total emission of the cathode is drawn to the anode, beyond which point the current is said to be *temperature limited.*

The relation between current and potential under conditions of space-charge limitation may be represented approximately by the equation

$$i_b = K_1 e_b^{3/2}$$

where K_1 is a constant depending on the physical dimensions of the tube.

Since the electron flow to the anode depends on the electrostatic field in the neighborhood of the cathode, the current may be controlled in part by the potential of another electrode in a position to influence this field. In the simple *triode* this additional electrode consists of a *grid*-like structure placed between cathode and anode. Because of its proximity to the cathode, the grid has more control of the field near the cathode than the anode has, but because of its open structure most of the electrons pass through to the anode. When the grid is operated at a negative potential with respect to the cathode, none of the electrons are taken by the grid.

The relation between anode current and grid and anode potentials is usually represented by the equation

$$i_b = K_2(e_b + \mu e_c)^{\eta} \tag{1}$$

where K_2 is a constant depending primarily on the cathode and grid dimensions, μ is the *amplification factor* of the tube (determined by the grid structure and grid-anode spacing), and η is usually between 1.5 and 2.5.

Other electrodes, usually grids, may be added to the structure, but the fundamental principles involved remain the same. For applications of high-vacuum tubes see Sections 7, 16, 17, 19, 20, and 21.

GAS-FILLED TUBES. In a *gas-filled diode*, the space contains sufficient gas at a low pressure to cause an appreciable fraction of the electrons passing between cathode and anode to collide with the gas molecules and thereby ionize them. The dislodged electrons pass on to the anode as additional anode current, while the positively charged ions are drawn to the cathode, but at a much lower velocity, owing to their greater mass, than the electrons possess. Because of this low velocity a given *positive-ion current* produces a much higher space-charge density than the same electron current. Therefore, the net space-charge density produced by the two currents (the algebraic sum of the positive and negative space charges) may approach zero, though the electron current is by far the larger.

Since the limitation of current in a high-vacuum diode is due to the negative space charge and this space charge is reduced by the positive ions, it follows that the presence of the gas increases the current which may flow at a given anode potential. Under normal conditions, the discharge is unstable at potentials much in excess of the ionizing potential of the gas, the current increasing until limited by the cathode emission. For this reason, the current is usually limited by an external resistance in series with the anode.

The relation between applied voltage and current may be expressed by the equation

$$i_b = \frac{e_{bb} - E_b}{r} \tag{2}$$

where E_b is the practically constant "anode drop" of the tube and r is the load resistance.

2. CLASSIFICATIONS

Vacuum tubes in general are classified according to many different structural and electrical characteristics.

Cathodes are *directly heated* if the actual emitter is also the resistance element which supplies the heat, and *indirectly heated* if the heat is supplied by conduction or radiation from a resistance element. The directly heated cathode is more efficient than the indirectly heated, but for many applications may not be heated by alternating current. The cathode *material* may be *tungsten, thoriated tungsten,* or *oxide-coated metal.* Tungsten is the least efficient and is generally used only in high-voltage tubes. Thoriated tungsten is much more efficient than tungsten and is used in many medium-voltage tubes (500 to 2000 volts anode potential) and some low-voltage tubes; at high voltages the emission life may be short. Neither of these materials is used for indirectly heated cathodes. Oxide-coated cathodes are the most efficient but (except in gas-filled tubes) are commonly used only in low-voltage tubes because of troubles from grid emission.

Anodes are *radiation cooled,* or *air cooled,* if no other cooling means is provided, or *water cooled* if provided with means for circulating water about or through the anode. Only high-voltage high-power (more than 6000 volts, 2500 watts rating) tubes are water cooled, except for special high-frequency tubes.

Rectifiers are *high vacuum* or *gas filled.* For high-power high-voltage applications, gas-filled tubes are usual, though high-vacuum tubes are used above 20,000 volts. For radio receivers both are used, gas-filled tubes predominating in other low-voltage applications.

Triodes are classified according to voltage and power rating, amplification factor, and special features such as designs for high-frequency oscillators, freedom from mechanical disturbance, etc. There are so many triodes of varying characteristics that almost any requirement can be met.

Multigrid tubes have been developed largely for specialized radio purposes, but the *screen-grid tubes,* including *suppressor-grid pentodes,* have wide application. In the screen-grid tube, the grid is screened or shielded from the anode, thus greatly reducing the feedback capacitance between plate and grid. For high-frequency voltage amplification, these tubes are used almost exclusively.

3. DEFINITIONS

Vacuum Tube. (Electron Tube.) A vacuum tube is a device consisting of an evacuated enclosure containing a number of electrodes between two or more of which conduction of electricity through the vacuum or contained gas may take place.

NOTE: The term is used in a more restricted sense to mean a device of this nature designed for such use as amplifier, rectifier, modulator, or oscillator.

Gas Tube. A gas tube is a vacuum tube in which the pressure of the contained gas or vapor is such as to affect substantially the electrical characteristics of the tube.

Mercury-vapor Tube. A mercury-vapor tube is a gas tube in which the active contained gas is mercury vapor.

High-vacuum Tube. A high-vacuum tube is a vacuum tube evacuated to such a degree that its electrical characteristics are essentially unaffected by gaseous ionization.

Thermionic Tube. A thermionic tube is a vacuum tube in which one of the electrodes is heated for the purpose of causing electron or ion emission.

Phototube. (Photoelectric Tube.) A phototube is a vacuum tube in which one of the electrodes is irradiated for the purpose of causing electron emission

Cathode-ray Tube. A cathode-ray tube is a vacuum tube in which the electron stream is directed along a confined path to produce non-electrical effects on the object upon which the electrons impinge.

NOTE: This classification includes cathode-ray oscillograph tubes, similar devices for television reception, electron microscopes, etc.

Cathode-ray Oscillograph Tube. A cathode-ray oscillograph tube is a cathode-ray tube in which the movement of an electron beam, deflected by means of applied electric and/or magnetic fields, indicates the instantaneous values of the actuating voltages and/or currents.

Diode. A diode is a two-electrode vacuum tube containing an anode and a cathode.

Triode. A triode is a three-electrode vacuum tube containing an anode, a cathode, and a control electrode.

Tetrode. A tetrode is a four-electrode vacuum tube containing an anode, a cathode, a control electrode, and one additional electrode ordinarily in the nature of a grid.

Pentode. A pentode is a five-electrode vacuum tube containing an anode, a cathode, a control electrode, and two additional electrodes ordinarily in the nature of grids.

Hexode. A hexode is a six-electrode vacuum tube containing an anode, a cathode, a control electrode, and three additional electrodes ordinarily in the nature of grids.

Heptode. A heptode is a seven-electrode vacuum tube containing an anode, a cathode, a control electrode, and four additional electrodes ordinarily in the nature of grids.

Octode. An octode is an eight-electrode vacuum tube containing an anode, a cathode, a control electrode, and five additional electrodes ordinarily in the nature of grids.

Multiple-unit Tube. A multiple-unit tube is a vacuum tube containing within one envelope two or more groups of electrodes associated with independent electron streams.

NOTE: A multiple-unit tube may be so indicated, as, for example; duodiode, duotriode, diode-pentode, duodiode-triode, duodiode-pentode, and triode-pentode.

Cathode. A cathode is an electrode which is the primary source of an electron stream.

Filament. A filament is a cathode of a thermionic tube, usually in the form of a wire or ribbon, to which heat may be supplied by passing current through it.

Indirectly Heated Cathode. (Equipotential Cathode, Unipotential Cathode.) An indirectly heated cathode is a cathode of a thermionic tube to which heat may be supplied by an independent heater element.

Heater. A heater is an electric heating element for supplying heat to an indirectly heated cathode.

Control Electrode. A control electrode is an electrode on which a voltage is impressed to vary the current flowing between two or more other electrodes.

Grid. A grid is an electrode having one or more openings through which electrons or ions may pass.

Space-charge Grid. A space-charge grid is a grid which is placed adjacent to the cathode and positively biased so as to reduce the limiting effect of space charge on the current through the tube.

Control Grid. A control grid is a grid, ordinarily placed between the cathode and the anode, for use as a control electrode.

Screen Grid. A screen grid is a grid placed between a control grid and an anode, and usually maintained at a fixed positive potential, for the purpose of reducing the electrostatic influence of the anode in the space between the screen grid and the cathode.

Suppressor Grid. A suppressor grid is a grid which is interposed between two electrodes (usually the screen grid and plate), both positive with respect to the cathode, in order to prevent the passing of secondary electrons from one to the other.

Anode. An anode is an electrode to which a principal electron stream flows.

Plate. Plate is a common name for the principal anode in a vacuum tube.

Electron Emission. Electron emission is the liberation of electrons from an electrode into the surrounding space. Quantitatively, it is the rate at which electrons are emitted from an electrode.

Thermionic Emission. Thermionic emission is electron or ion emission due directly to the temperature of the emitter.

Secondary Emission. Secondary emission is electron emission due directly to the impact of electrons or ions.

Grid Emission. Grid emission is electron or ion emission from a grid.

Emission Characteristic. An emission characteristic is a relation, usually shown by a graph, between the emission and a factor controlling the emission (as temperature, voltage, or current of the filament or heater).

Cathode Current. Cathode current is the total current passing to or from the cathode through the vacuous space.

NOTE: This term should be carefully distinguished from heater current and filament current.

Filament Current. Filament current is the current supplied to a filament to heat it.

Filament Voltage. Filament voltage is the voltage between the terminals of a filament.

Heater Current. Heater current is the current flowing through a heater.

Heater Voltage. Heater voltage is the voltage between the terminals of a heater.

Grid Current. Grid current is the current passing from or to a grid through the vacuous space.

Grid Voltage. Grid voltage is the voltage between a grid and a specified point of the cathode.

Grid Bias. Grid bias is the direct component of grid voltage.

Grid Driving Power. Grid driving power is the integral of the product of the instantaneous values of the alternating components of the grid current and voltage over a complete cycle.

Anode Current. (Plate Current.) Anode current is the current passing to or from an anode through the vacuous space.

Anode Voltage. (Plate Voltage.) Anode voltage is the voltage between an anode and a specified point of the cathode.

Peak (or Crest) Forward Anode Voltage. Peak (or crest) forward anode voltage is the maximum instantaneous anode voltage in the direction in which the tube is designed to pass current.

Peak (or Crest) Inverse Anode Voltage. The peak (or crest) inverse anode voltage is the maximum instantaneous anode voltage in the direction opposite to that in which the tube is designed to pass current.

Tube Voltage Drop. Tube voltage drop in a gas- or vapor-filled tube is the anode voltage during the conducting period.

Anode Dissipation. Anode dissipation is the power dissipated in the form of heat by an anode as a result of electron and/or ion bombardment.

Gas Current. A gas current is a current flowing to an electrode and composed of positive ions which have been produced as a result of gas ionization by an electron current flowing between other electrodes.

Leakage Current. A leakage current is a current which flows between two or more electrodes by any other path than across the vacuous space.

Electrode Conductance. (Variational.) Electrode conductance is the ratio of the in-phase component of the electrode alternating current to the electrode alternating voltage, all other electrode voltages being maintained constant.

NOTE: As most precisely used, the term refers to infinitesimal amplitudes.

Electrode Resistance. (Variational.) Electrode resistance is the reciprocal of the electrode conductance.

Electrode Admittance. Electrode admittance is the ratio of the alternating component of the electrode current to the alternating component of the electrode voltage, all other electrode voltages being maintained constant.

NOTE: As most precisely used, the term refers to infinitesimal amplitudes.

Electrode Impedance. Electrode impedance is the reciprocal of the electrode admittance.

Transadmittance. Transadmittance between two electrodes is the ratio of the alternating component of the current of one electrode to the alternating component of the voltage of the other electrode, all other electrode voltages being maintained constant.

NOTE: As most precisely used, the term refers to infinitesimal amplitudes.

Transconductance. Transconductance between two electrodes is the ratio of the in-phase component of the alternating current of one electrode to the alternating voltage of the other electrode, all other electrode voltages being maintained constant.

NOTE: As most precisely used, the term refers to infinitesimal amplitudes.

Control-grid—Plate Transconductance. (Transconductance, Mutual Conductance.) Control-grid—plate transconductance is the name for the plate-current to control-grid voltage transconductance.

Conversion Transconductance. Conversion transconductance is the ratio of the magnitude of a single beat-frequency component $(f_1 + f_2)$ or $(f_1 - f_2)$, of the output electrode current to the magnitude of the control electrode voltage of frequency f_1 under the conditions that all direct electrode voltages and the magnitude of the electrode alternating voltage f_2 remain constant.

NOTE: As most precisely used, the term refers to an infinitesimal magnitude of the voltage of frequency f_1.

Mu Factor. The mu factor is the ratio of the change in one electrode voltage to the change in another electrode voltage, under the conditions that a specified current remains unchanged and that all other electrode voltages are maintained constant. It is a measure of the relative effect of the voltages of two electrodes on the current in the circuit of any specified electrode.

NOTE: As most precisely used, the term refers to infinitesimal changes.

Amplification Factor. The amplification factor is the ratio of the change in plate voltage to a change in control-electrode voltage, under the conditions that the plate current remains unchanged and that all other electrode voltages are maintained constant. It is a measure of the effectiveness of the control-electrode voltage relative to that of the plate voltage on the plate current. The sense is usually taken as positive when the voltages are changed in opposite directions.

NOTE: As most precisely used, the term refers to infinitesimal changes. *See also* Mu Factor.

Electrode Characteristic. An electrode characteristic is a relation, usually shown by a graph, between an electrode voltage and current, other electrode voltages being maintained constant.

Transfer Characteristic. A transfer characteristic is a relation, usually shown by a graph, between one electrode voltage and another electrode current.

Interelectrode Capacitance. Interelectrode capacitance is the direct capacitance between two electrodes.

Electrode Capacitance. Electrode capacitance is the capacitance of one electrode to all other electrodes connected together.

Input Capacitance. The input capacitance of a vacuum tube is the sum of the direct capacitances between the control grid and cathode and such other electrodes as are operated at the alternating potential of the cathode.

Output Capacitance. The output capacitance of a vacuum tube is the sum of the direct capacitances between the output electrode (usually the plate) and the cathode and such other electrodes as are operated at the alternating potential of the cathode.

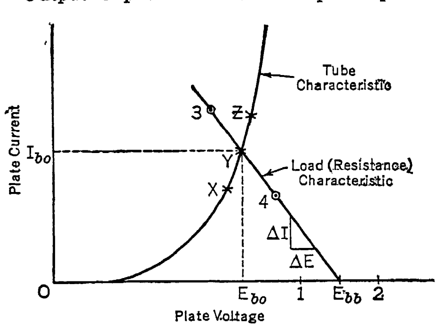


FIG. 1. Tube and Load Characteristics

QUIESCENT POINT AND OPERATING RANGE. The *quiescent point* is the point on the plate characteristic that represents operating conditions with no signal applied to the grid. With a load in the plate circuit, it may be determined as follows. For a load resistance of r ohms the slope of the load characteristic line is $1/r = -\Delta I/\Delta E$, drawn from the point E_{bb} (Fig. 1) corresponding to the plate supply voltage. The intersection y of the two characteristic curves is the quiescent operating point, giving a plate voltage of E_{bo} and a plate current of I_{bo}.

The plate supply voltage E_{bb} is divided into the voltage drop across the tube, E_{bo}, and the voltage drop across the load, $E_{bb} - E_{bo}$.

When the plate supply voltage is varied over the range 1 to 2, the load resistance line will shift parallel to the position shown and over the *operating range* 1 to 2. The intersections of the load line with the tube characteristic determine the corresponding variation in E_{bo} and I_{bo}.

When the plate supply voltage is constant but the tube characteristic is changed, for example by changing the grid voltage, the intersection of the characteristics for the changed values of grid voltage with the load line determines the change in I_{bo} and E_{bo}.

In Fig. 1 a voltage in the grid circuit changes the characteristic curve so that it intersects the load line at points 3 or 4. The projections of these points on the two axes show the change in I_{bo} and E_{bo}.

The tube characteristic, unlike the load characteristic, cannot be shifted parallel to the initial position to represent other operating conditions, since it usually changes shape. The tube characteristics should be known for a few voltages throughout the range of operation. Intermediate values may be interpolated.

A-C EQUIVALENT CIRCUIT. In the circuit of Fig. 2 the a-c voltage E_g in the grid circuit produces an a-c plate current I_p and an a-c plate voltage E_p. For small a-c voltages

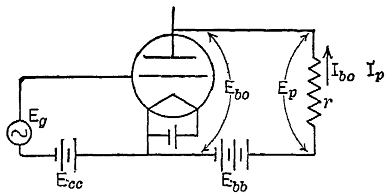


FIG. 2. Triode with Resistance Load

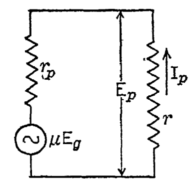


FIG. 3. Equivalent A-c Circuit of Tube with Resistance Load

the tube is equivalent to a generator with an internally generated voltage μE_g and internal resistance r_p. Figure 3 is the *a-c equivalent* of the circuit in Fig. 2.

From Fig. 3 the *a-c plate current* is

$$I_p = \frac{\mu E_g}{r_p + r} \tag{3}$$

The *a-c plate voltage* is

$$E_p = I_p r = \mu E_g \frac{r}{r_p + r} \tag{4}$$

The *voltage amplification* is

$$A = \frac{E_p}{E_g} = \frac{\mu r}{r_p + r} \tag{5}$$

The *power output* of the tube is

$$P_o = E_p \cdot I_p = \frac{(\mu E_g)^2 r}{(r_p + r)^2} \tag{6}$$

Figure 4 illustrates an a-c equivalent of the circuit of Fig. 2 in which the tube is represented as a *generator of constant current* $I = g_m E_g$.

The *a-c plate voltage* is

$$E_p = I \frac{r_p r}{r_p + r} = (g_m E_g) \frac{r_p r}{r_p + r} \tag{7}$$

The *a-c plate current* is

$$I_p = \frac{E_p}{r} = g_m E_g \left(\frac{r_p}{r_p + r}\right) \tag{8}$$

The constant-current form of representation is convenient for calculation when the load consists of a number of parallel elements or when the plate resistance of the tube is high and $r_p/(r_p + r)$ approaches unity.

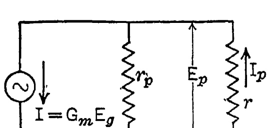


FIG. 4. Equivalent A-c Circuit of Tube with Resistance Load.

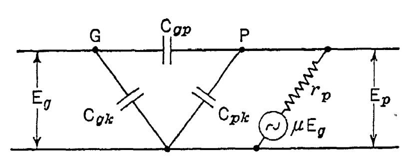


FIG. 5. Equivalent Circuit of Triode

When the tube is amplifying a-c voltages at frequencies at which the *capacitances of tube electrodes, socket, and wiring are not negligible* the equivalent a-c circuit is as shown in Fig. 5.

The capacitances marked C_{gp}, C_{gk}, C_{pk} represent the grid-plate, grid-cathode, and plate-cathode capacitances (see Section 5, Article 25). Any socket and wiring capacitances can be added in parallel. In screen-grid and other multielectrode tubes where the additional electrodes are grounded the circuit reduces to an equivalent-triode network similar to Fig. 3.

BALLAST TUBES AND VOLTAGE REGULATORS. A *ballast tube* is used as a series resistance to limit the load current. It is designed for a definite current and voltage drop. Over the useful portion of its characteristic a large change in voltage accompanies a small change in current. As a result of this characteristic a large part of any line-voltage change is absorbed by the ballast tube and a relatively small change occurs at the load.

In reading data for the characteristic curves, it may be necessary to allow a few minutes after each change in voltage for the tube to reach its temperature equilibrium condition.

The performance may be determined graphically by a method similar to that described for determining the quiescent operating point by plotting the line for a resistance load with the experimentally determined current-voltage characteristic of the ballast tube.

A *voltage-regulator tube* is operated in parallel with the load. Its characteristic shows a large change in current for a small change in voltage near the operating point. When connected in parallel with the load with a suitable resistance effective in the supply voltage source any change in the supply voltage will cause a change in the current in the voltage-regulator tube such that the voltage across the regulator tube and load remains practically constant.

The voltage-regulator tube is designed for a definite voltage and is operated between specified minimum and maximum current limits. A certain *starting voltage* somewhat higher than the operating voltage is required.

The operating characteristics are most conveniently obtained in a normal operating circuit.

4. METHODS OF MEASURING TUBE CURRENTS AND PARAMETERS

The characteristic relations between the direct voltages and currents of the electrodes of a tube may be obtained in a *static-characteristic measuring circuit* arranged as in Fig. 6.

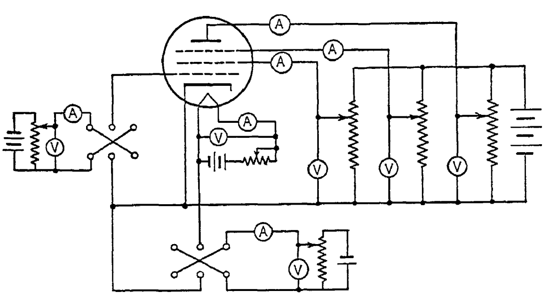


FIG. 6. Circuit for Measuring Static Characteristics

The voltages applied to the different electrodes as illustrated are measured from a *unipotential cathode.* If the tube has a *filamentary cathode* it is understood that, when operating with *direct-current filament supply* such as when measuring static characteristics, the electrode voltages are measured from the negative filament terminal. With *alternating-current* operation of a filamentary cathode, the center of the filament is used as the datum of potential and the electrode voltages are corrected for one-half of the filament voltage. Ordinarily only the control-grid bias voltage is made more negative by approximately one-half of the peak alternating voltage on the filament.

FILAMENT OR HEATER CHARACTERISTIC. The *filament* or *heater current* is obtained for several values of filament or heater voltage ranging from values producing temperatures too low to give appreciable electron emission to values producing the maximum safe operating temperature. The other electrodes should be at zero voltage.

The curves are plotted with filament or heater voltage as abscissas and filament or heater current and power as ordinates.

CATHODE HEATING TIME. The *cathode heating time* is defined for purposes of measurement as the interval from the time of application of filament or heater voltage to the time at which the rate of increase of plate current is a maximum.

If the primary winding of a transformer is connected in the plate circuit and a meter is connected to the secondary winding, then the instant at which the rate of increase of plate current is a maximum is indicated by a maximum reading on the meter.

The voltage at the terminals of the filament or heater should remain constant at the rated or specified value.

Because of the very slow rate of change of plate current in the usual tube, the characteristics of the transformer or meter do not greatly influence the result. A step-down transformer and a current meter which is sufficiently damped though not too sluggish are suitable.

EMISSION CHARACTERISTIC. The *emission characteristic* shows the emission current plotted as a function of the cathode heating power.

The readings are obtained with all electrodes, except the cathode, connected together as the anode and with sufficient positive voltage applied to the electrodes to draw the entire emission current from the cathode. Since the emission current at normal filament power may be so great as to damage the tube, readings are taken at lower filament powers

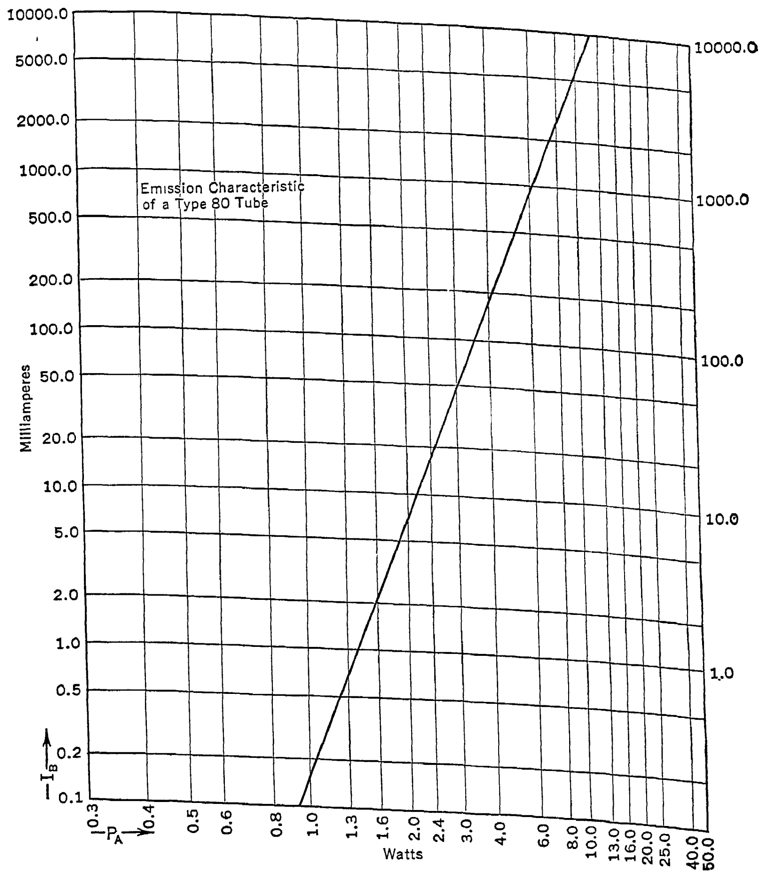


FIG. 7. Emission Characteristic

only, and normal emission current is obtained by *extrapolation.* A suitable procedure is as follows, the values applying to ordinary receiving tubes. Readings of cathode heating power are taken with emission currents of 0.1, 0.2, 0.5, 1.0, 2.0, and 5.0 ma, with 45 volts positive applied to the anode. The results are plotted in *Davisson coordinates* (see Fig. 7), which are a special system of curvilinear coordinates. If the emission follows *Richardson's temperature law* and the cathode cooling follows the *Stefan-Boltzmann law* of radiation, the characteristic will be a straight line when plotted in these coordinates. The observed points may be extended or extrapolated to obtain the emission at normal filament power.

The emission characteristic for a type-80 tube plotted according to the above procedure is shown in Fig. 7.

If the curve of the experimental data plotted in Davisson coordinates is *not a straight line* this may be due to one or more of the following conditions:

1. Departure from the Stefan-Boltzmann cooling (bends downward).
2. Anode voltage too low to draw off all the electrons (bends downward).

3. Effect of cooling due to heat of evaporation of electrons (bends downward). The cooling due to electron evaporation amounts to approximately ϕI_s watts, where I_s represents the emission current in amperes and ϕ represents the work function of the cathode in volts. This effect may be considerable in transmitting tubes where the currents are high and in tungsten-filament tubes where the work function is large.

4. Poor vacuum (gas ionization effects) (bends upward).

5. Heating of the anode by the emission current (bends upward).

6. Progressive change in activity of the cathode.

Reliable analytical data cannot be obtained by this method when these extraneous effects are appreciable.

ELECTRON EMISSION. Normal *electron emission* is determined with the filament voltage adjusted to the normal rated value.

All electrodes in the tube, except the cathode and heater, are connected together, and a sufficiently positive voltage with respect to the cathode is applied to them to obtain practically the full electron emission.

For *power-type tubes* this test is not advisable on account of possible damage to the tube. The method of extrapolation described under Emission Characteristic should be used.

For *receiving-type tubes* a check on the emissive condition of the cathode can usually be made safely if the time of application of the voltage is not permitted to exceed that required for rapid reading of the emission current meter. An anode voltage of about 45 volts is used.

Since this test usually results in the liberation of gas and abnormal heating of the electrodes, it should be postponed until after the completion of other tests, or a sufficient time should elapse between this and other tests for clean-up and return to normal temperature conditions.

For tubes with extremely low cathode heating power, such as the oxide-coated low-filament-current types 1R5, 1S4, 1T4, etc., this test is neither reliable nor safe for the tube. In *checking the emission* of tubes of this type a low filament voltage is applied and gradually increased until a specified emission (less than normal for the particular type of tube, usually 3 to 5 ma) is obtained. The filament voltage required to obtain the specified emission is an indirect measure of the cathode or filament activity. This is an arbitrary method suitable for comparing tubes of the same type.

In general a *safe method for reading emission* under all different conditions consists in using a rotating contactor to apply the voltage for only a small fraction of the time. An oscillograph is used to read the emission current which flows during the small interval of time that the voltage is applied.

GRID CHARACTERISTIC. A *grid characteristic curve* shows the current in a grid electrode as a function of the voltage on this electrode. The voltage on all the remaining electrodes is held constant. A *family of curves* is obtained by using a different value of voltage on one of the remaining electrodes for each curve.

The grid characteristic used most frequently is that of the control grid with the plate voltage as the parameter for the several curves.

In reading data for the curves the current should not be allowed to flow long enough to cause abnormal heating of the grid. The readings should be taken near enough together to show any irregularities due to secondary emission or gas.

IONIZATION, LEAKAGE, AND STRAY EMISSION CURRENTS. In vacuum tubes the normally small currents due to *ionization*, *leakage*, and *electron emission* from electrodes other than the cathode, although usually negligible in the plate and other current-carrying electrodes, may have an appreciable effect in the control-grid circuit of the tube.

The total current flowing to the negatively biased control grid may be divided into components as follows:

1. Electrons from the cathode which reach the grid by virtue of *contact potential and initial velocities*.
2. Electrons from other electrodes to the control grid.
3. *Ionization current.*
4. *Leakage current.*
5. *Electron emission* from the control grid.

Figure 8 illustrates the contributions of the various sources enumerated with the exception of 2, which is generally negligible.

The several components may be separated and measured by the following methods.

The *leakage current* is measured with a direct voltage applied between any two electrodes and without any connections on the other electrodes. The tube should be operated with normal voltages and currents until all parts have reached full operating temperature. The filament voltage is then disconnected and the leakage currents read while the insulation is at normal operating temperature. The tube should be complete with its base but without socket or holder. The test voltage should be specified. Normal maximum operating voltage is preferable.

If any of the electrodes remain hot enough to emit electrons an error will be introduced into the leakage readings.

The *grid emission* can be measured by noting the current at a bias sufficiently negative (point A in the figure) to stop the plate current, since at this point (A) the ionization current (3), being proportional to the plate current, is negligible.

The grid emission is found by subtracting the leakage current from the grid current at the point A. If the leakage current is negligible the test gives the grid emission directly.

A *direct measurement of grid emission* can be made (when leakage current is negligible) by connecting the test voltage between the grid and plate without any connection to the cathode. The positive voltage on the plate draws the electrons emitted by the grid. The cathode should be at its normal operating temperature. The tube should be operated with normal operating voltages for a time preceding this test, and then quickly switched to the emission test circuit and grid emission noted while the electrodes are still approximately at their normal temperatures.

The *ionization current* (3) is the difference between the total grid current and the sum of the leakage (4) and emission (5) currents in the range of grid bias over which this difference is proportional to the plate current. Departure from this proportionality indicates the start of electron current (1) to the grid.

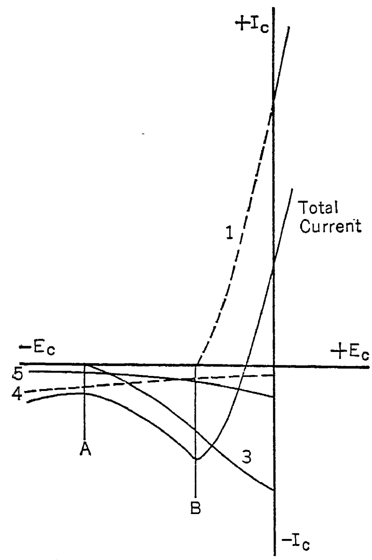


FIG. 8. Stray Electron Currents

PLATE CHARACTERISTIC. The *plate characteristic* gives the plate current as a function of the plate voltage, the voltages on the other electrodes being held constant. A family of curves may be obtained by using a series of voltages on one of the other electrodes. A series of control-grid voltages is ordinarily used for the different curves. For examples see Figs. 21, 24, and 27.

The data for the curves are read in the static characteristic test circuit. For the range of currents and voltages beyond the normal average values, it is sometimes necessary to employ a method which is rapid enough to avoid heating of the electrodes. When a well-regulated voltage source is available the voltages may be set to the desired values and a switch closed only long enough for rapidly reading the current on a suitably damped meter.

GRID-PLATE CHARACTERISTIC. The *grid-plate characteristic* or *transfer characteristic* gives the plate current vs. grid voltage for the condition of constant voltage on the remaining electrodes. Such a transfer characteristic may be taken for any of the grids of a multigrid tube. The curves generally used show the control-grid voltage as abscissa and plate current as ordinate, several curves being plotted each for a different plate voltage. See for examples Figs. 22, 25, and 29.

The data for the curves are obtained in the static-characteristic test circuit, the same precautions being observed as in reading data for plate characteristics.

CONDUCTANCE. The *conductance* of an electrode may be obtained from the characteristic curve showing the electrode current vs. the electrode voltage. The slope of this curve at any point gives the electrode conductance at the voltages represented by the point. The accuracy of the measurement as determined in this way from the static characteristics may be made as good as desired by reading small current and voltage increments with sufficient accuracy.

The *electrode resistance* is the reciprocal of the conductance. For example, the *plate resistance* is the reciprocal of the plate conductance, the *grid resistance* is the reciprocal of the grid conductance, etc.

When many readings are to be made, a method of *direct measurement* is most convenient. One means might be to read the alternating current produced by a small alternating voltage in the circuit. In general a *Wheatstone bridge circuit* is preferred as shown in Fig. 9.

In this circuit a small alternating voltage (about 0.5 volt, 1000 cps) is applied to terminals 1–3 of the bridge. The electrodes being measured are connected to terminals 1–2, and the bridge is balanced with r_2 for minimum sound in the telephone receivers. The *conductance* of the electrode j is given by

$$g_j = \frac{r_2}{r_1 r_3} \tag{9}$$

If the receivers are preceded by an amplifier the alternating input voltage may be kept small enough so that the measurements are independent of the magnitude of this voltage.

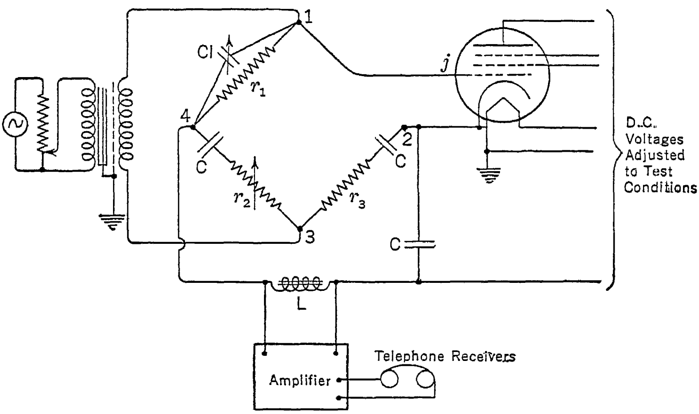


FIG. 9. Conductance Measurement Circuit

This also reduces any harmonics produced by non-linearity of the characteristic. The *stray capacitances* of the wiring and parts of the bridge should be kept low or balanced by a suitable arrangement of parts. A variable capacitance C_1 connected across r_1 can be used to balance different *tube capacitances.*

The *voltage drop in the bridge* circuit should be taken into account in determining the electrode voltage of the tube. The inductance L provides a low d-c resistance path across the amplifier. The blocking condensers C keep the direct current in a path of constant resistance which simplifies the correction.

TRANSCONDUCTANCE. The *transconductance* is a measure of the change in current in one electrode produced by a voltage on another electrode. In a tube having several electrodes the transconductance may be measured between any two electrode circuits,

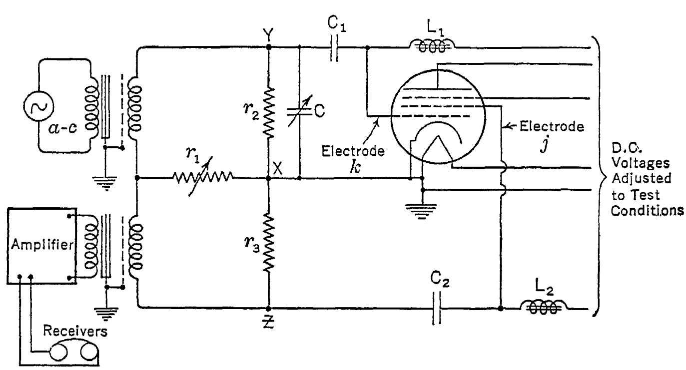


FIG. 10. Transconductance Measurement Circuit

the cathode being common to both circuits. For example, the grid-plate transconductance, or *mutual conductance,* may be determined from the curves for the grid-plate transfer characteristic since this shows the plate current vs. grid voltage. The slope of this curve at a certain point gives the transconductance at the voltages represented by the point.

The transconductance may be *measured directly* by means of the circuit of Fig. 10. An alternating voltage as small as convenient is applied in this circuit, and the resistor r_1 is adjusted for minimum sound in the receivers. The stray capacitances due to the wiring and the tube may be balanced by means of the condenser C connected either between points X–Y or X–Z as determined by trial. The direct voltages are supplied to the electrodes through the choke coils L_1 and L_2. The reactance of the choke coils at the frequency of the applied alternating voltage should be large with respect to r_2 and r_3. If the resistances of the choke coils cause an appreciable drop in direct voltage the electrode voltages should be corrected accordingly.

At balance the *transconductance* is given by

$$g_{jk} = \frac{r_1}{r_2 r_3} \tag{10}$$

This relation assumes that the *electrode conductance* is negligible. If the conductance of electrode j is $1/r_j$ and that of k is $1/r_k$ the transconductance is given by

$$g_{jk} = \frac{r_1}{r_2 r_3} \frac{(r_k + r_2)}{r_k} \cdot \frac{(r_j + r_3)}{r_j} \tag{11}$$

The following circuit constants will cover a range of transconductance from 1 micromho to 10,000 micromhos.

$r_1 = 10 \times 0.1$ ohm		$r_2 = 1000$ ohms
$10 \times 1.$ ohm		
$10 \times 10.$ ohms		$r_3 = 100$ ohms
$10 \times 100.$ ohms		

With L_1 equal to 50 henrys and L_2 equal to 10 henrys and a 1000 cps alternating voltage the error in using the simplified equation will be less than 2 per cent if r_j and r_k are greater than 10,000 ohms and 100,000 ohms, respectively.

MU FACTOR. The *mu factor* is the ratio of the voltages in two electrode circuits required to maintain constant current in the circuit of any specified electrode. It may be determined from the static characteristic curves or measured by a balance method. For example, the *amplification factor* is a special case in which the control-grid voltage and plate voltage are changed in such a way as to maintain constant plate current.

In the circuit of Fig. 11 the electrode in which the current is to be held constant is connected to point A. The other two electrodes entering directly in the measurement

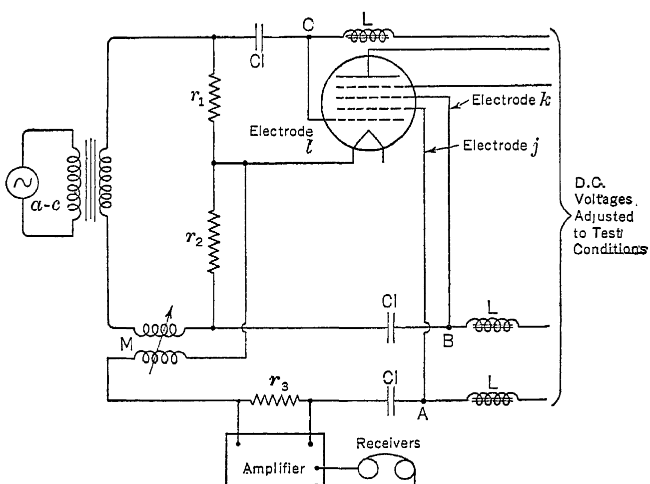


FIG. 11. Mu Factor Measurement Circuit

are connected to points B and C. When r_1 and r_2 and M are adjusted for minimum sound in the telephone receivers, the *mu factor* is given by

$$\mu_{jkl} = \frac{r_2}{r_1}$$

The direct voltages to the electrodes under test may be supplied through choke coils L having a reactance at the frequency of the alternating voltage which is high with respect to the resistances r_1, r_2, and r_3. The resistances of the chokes should be low enough to cause negligible loss in direct voltage or the true electrode voltage determined by subtracting the loss. If the *electrode conductances* are not negligible in comparison with r_1, r_2, and r_3 a correction should be made for this.

INTERELECTRODE CAPACITANCE. The *capacitance between the electrodes* of a tube may be measured in various ways. It is preferable to read the *direct capacitance* between any two electrodes rather than the *total capacitance* between an electrode and all other electrodes. The readings are normally made with the tube *cold* and no direct voltages applied. When the tube is *heated* the capacitance changes a small amount owing to the presence of space charge, but the change is ordinarily negligible. The tube should be complete with base, though the socket capacitance is not included in the measurement. For most reliable results, readings should not be taken with any electrodes disconnected, and the tube should be mounted in a specified way with respect to any shields. For *indirectly heated types* the filament and cathode should be tied together.

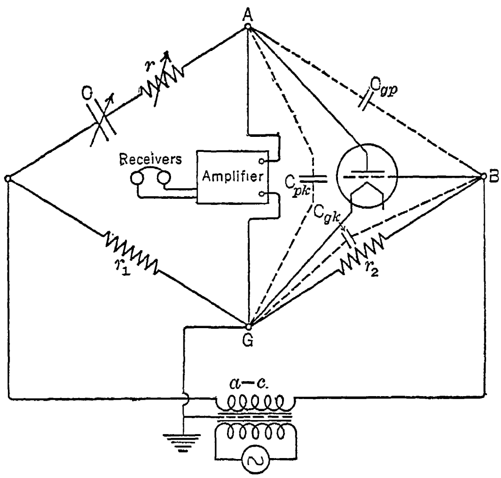


FIG. 12. Electrode Capacitance Measurement Circuit

A *bridge method* for the measurement of *direct interelectrode capacitance* in a *triode* is shown in Fig. 12. The capacitance to be measured is connected to the terminals *A–B*. The figure shows the grid-plate capacitance C_{gp} connected for measurement. The effect of the grid-cathode capacitance C_{gk} in the circuit across r_2 is ordinarily negligible owing to the low resistance of r_2. The plate-cathode capacitance C_{pk} is across the amplifier and telephone receivers, which does not affect the balance. The standard capacitance C and resistance r are adjusted for minimum sound in the receivers. When the bridge is balanced the capacitance is

$$C_x \equiv C_{gp} = \frac{r_1 C}{r_2} \tag{12}$$

An error in the reading may result if appreciable *leakage resistance* exists across C_x.

5. VACUUM-TUBE OPERATION

ELECTRON TRANSIT TIMES AND INERTIA EFFECT. Since the single electron has a mass of 9.035×10^{-28} gram there will be a force acting on the electron in the presence of an electric field which is proportional to the product of the field strength and the electron unit of charge. This force will accelerate the electron, giving it finite velocities. If the electron has started from rest and attains final velocities much smaller than the speed of light, the velocity in practical units may be expressed as

$$v = 5.95 \times 10^7 \sqrt{V} \quad \text{cm per sec}$$

where V is in volts. It is common practice to express the velocity of an electron in terms of the voltage in the above expression instead of the usual centimeters per second.

If, however, the velocity of the electron has been accelerated to velocities not negligible compared to the speed of light, c, which is 3×10^{10} cm per sec, the mass of the electron, m, will increase in the ratio

$$\frac{m}{m_e} = 1.96 \times 10^{-6}V + [(1.96 \times 10^{-6}V)^2 + 1]^{1/2} \tag{13}$$

where m_e is the mass of the electron at rest. At 100,000 volts the increase in mass is about 22 per cent, whereas at 1000 volts the increase is only about 0.2 per cent. At very high voltages, therefore, the velocity of an electron will be less than the above expression by the square root of m_e/m, the resulting velocity being

$$v = 5.95 \times 10^7 \sqrt{V \frac{m_e}{m}} \quad \text{cm per sec} \tag{14}$$

The transit time of an electron becomes of importance at very high frequencies and as would be expected depends upon the electron velocity expressed at some convenient point and the distance it has traveled. The constant K in the following expressions is a factor taking into account the geometry of the electric field and the effect of space charge, if present, on the electric field. For all the cases where the electron has been assumed to start with zero velocity the transit time, τ, is

$$\tau = \frac{KD}{5.95 \times 10^7} \times \frac{1}{\sqrt{V}} \quad \text{sec} \tag{15}$$

where D, in centimeters, is the distance traveled from rest to the point of voltage V, in volts. The dimensionless constant K has the following values: $K = 1$ for a field-free space; $K = 2$ for field between parallel planes without space charge; $K = 3$ for field between parallel planes with space charge.

For concentric cylinders the values of K are given in Fig. 13. The A and B curves are for the usual case where the higher voltage is on the outside cylinder (the cathode being

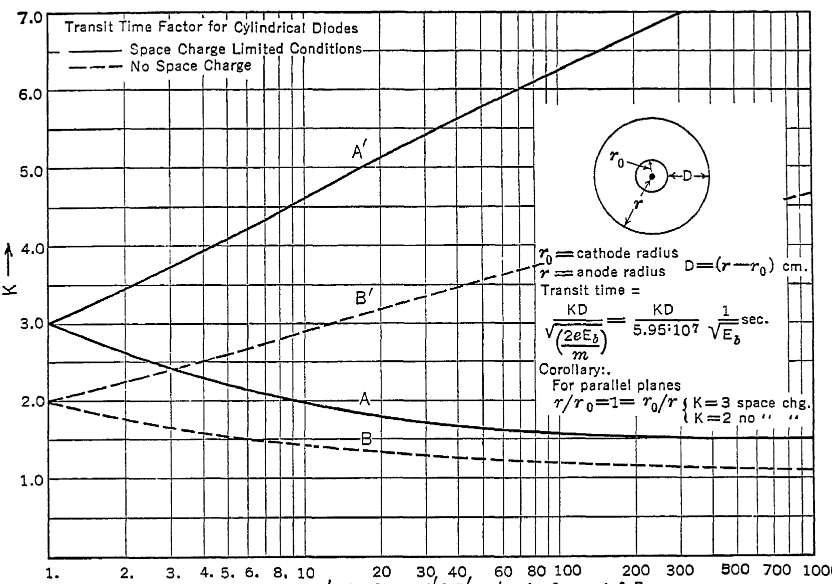


FIG. 13. Electron-transit Time for an Electron Starting with Zero Initial Velocity. Chart Includes Cylindrical and Parallel Plane Structures in an Electric Field with and without Space Charge. (Courtesy *RCA Review*.)

the inner one); the A' and B' curves are for the case where the positive voltage is on the inner cylinder.

In many cases one is interested in the transit time when the electron is not starting from rest, as, for example, for the electron transit from the control grid to the screen grid region of a pentode. For this case

$$\tau = \frac{KD}{5.95 \times 10^7} \times \frac{1}{\sqrt{V_1}} \frac{1}{\sqrt{V_2/V_1} + 1} \quad \text{sec} \tag{16}$$

where K and D are defined as before, V_1 and V_2 are the voltages of the two points between which the electron is passing, and $V_2 > V_1$.

THE INPUT ADMITTANCE OF VACUUM TUBES. The input admittance of negative-grid-controlled vacuum tubes may have deleterious effects especially at high frequencies. Its conductive component is, to a first order, proportional to the square of the input-signal frequency. Being a power-absorbing element it will thus broaden the frequency-response characteristic of the input circuit and lower the anticipated gain of the preceding stage. Furthermore, it also varies with the tube transconductance or tube gain. The reactive component of admittance is fortunately independent of frequency, but unfortunately it varies with the tube transconductance. For a triode or a pentode radio-frequency amplifier the effect is to increase the input capacitance of the tube in the order of 1 to 2 $\mu\mu$f as the tube gain is varied from cutoff to its maximum value. At low frequencies this small capacitance change is made negligible by padding the input circuit with a sufficiently large fixed capacitance, but at high frequencies this precaution is impractical. These admittance variations which change with the tube transconductance may be greatly reduced by the use of an unbypassed cathode resistor so positioned that it will be common to both the input and output circuits of the tube.

The input conductance, g_g, of a vacuum tube consists essentially of two components. One of these components, g_t, has its origin in the electron-transit time phenomena, and the other, g_L, has its origin in that portion of the cathode-lead inductance which is common to both the input and the output terminals of the vacuum tube. Quantitatively the transit-time loading, g_t, as determined by D. O. North is

$$g_t = \frac{4\pi^2 f^2 \tau_1^2 g_{1k}}{180}\left\{9 + 44\frac{\tau_2}{\tau_1} + 45\left(\frac{\tau_2}{\tau_1}\right)^2 - 2\frac{\tau_2}{\tau_1}\frac{17 + 35\,(\tau_2/\tau_1)}{1 + (v_p/v_g)} + \frac{20\,(\tau_2/\tau_1)^2}{[1 + (v_p/v_g)]^2}\right\} + \cdots \quad (17)$$

where g_t is the transit time input loading in mhos, g_{1k} is the signal-grid-to-cathode transconductance in mhos, f is the signal frequency in cycles per second, τ_1 is the transit time in seconds for an electron to move from the cathode to the plane of the signal grid, τ_2 is the transit time in seconds for an electron to move from the signal grid plane to the plate, and v_p and v_g are the electron d-c velocities at the plate and grid respectively. It should be noted that for a well-screened tetrode or pentode the transit time, τ_2, and the plate velocity, v_p, are with reference to the screen grid. Also it should be noted that the grid-to-cathode transconductance of a triode is the same as its grid-to-plate transconductance, whereas for a pentode or tetrode the grid-to-cathode transconductance has a value equal to the grid-to-plate transconductance multiplied by the ratio of cathode current to plate current. The last two terms in the brackets of the above expression are only of second order of magnitude in the conventional amplifier tubes since the ratio v_p/v_g is approximately equal to 10. Inspection of this expression further indicates that considerable loading may be contributed to the input circuit by electrons moving in the space between the control grid and plate (or screen grid). As a precaution in using the above relationship, the simplifying assumptions applied for its derivation are listed below.

1. The transit times, τ_1 and τ_2, are small compared to the period, $1/f$.
2. The electrodes are parallel planes.
3. The initial velocity of the emitted electrons is zero and the emission is ample, so that the three-halves power of voltage vs. current holds in the cathode-grid region.
4. The amplification factor of the signal grid is high, so that electrons on one side do not appreciably influence the field on the other side.
5. The grid is an equipotential plane surface.
6. The alternating voltage at the grid is very small with respect to the effective static grid potential.
7. The alternating voltage at the plate (or screen grid) is zero.
8. The potential between grid and plate (or screen grid) is substantially linear and free of space charge.

That part of the input conductance, g_L, which stems from the cathode-lead inductance may be expressed to a first approximation, as shown by J. O. Strutt and Van der Ziel, by

$$g_L = 4\pi^2 f^2 L_k C_{gk} g_{1k} \quad (18)$$

where g_L is the input loading due to cathode-lead inductance in mhos, f is the signal frequency in cycles per second, L_k is the cathode-lead inductance common to both input and output circuits in henrys, C_{gk} is the capacitance between the signal grid and the cathode in farads, and g_{1k} is the signal-grid-to-cathode transconductance in mhos.

Since both the aforementioned components of input loading are proportional to the signal frequency squared and to the tube transconductance, the two effects are therefore impractical to measure separately. It is also difficult to measure the cathode-lead inductance. If one estimates this inductance to be 5 to 10 millimicrohenrys for the glass miniature tubes and 10 to 15 millimicrohenrys for the single-ended metal or loctal tubes where the lower values are used for those tubes having two cathode leads, we find that from 10 per cent to about 35 per cent of the total input loading is contributed by the effect of cathode-lead inductance.

The screen-grid lead inductance, L_{g2}, in a pentode amplifier or tetrode amplifier introduces negative input loading, g_{L2}, which quantitatively is

$$g_{L2} = -4\pi^2 f^2 L_{g2} C_{g1g2} g_{12} \tag{19}$$

where f is the frequency in cycles per second, L_{g2} is the screen-grid lead inductance in henrys, C_{g1g2} is the capacitance between the signal grid and the screen grid in farads, and

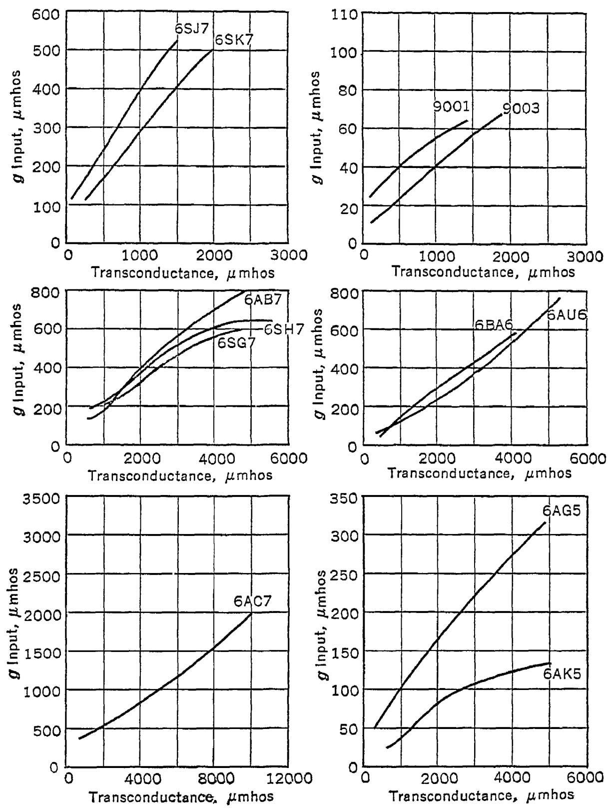


FIG. 14. Input-loading Conductance vs. Grid-plate Transconductance for a Group of R-f Pentode Amplifying Tubes at 100 Mc.

g_{12} is the transconductance from signal to screen grid. This negative loading is small in magnitude compared to the positive loading introduced by the cathode-lead inductance, and for the typical pentode amplifier ($g_{12} \approx 0.2g_{1k}$, $C_{g1g2} < C_{gk}$, and $L_{g2} \approx L_k$) it will reduce the total loading by only a few per cent. If, however, additional inductance is placed in series with the screen lead it should be possible to neutralize the positive input loading completely, with the distinct disadvantage that instability or even parasitic oscillations may be produced in the amplifier stage.

If the static tube voltages are held constant so that the electron-transit times do not vary and the frequency only is varied, we may write the total input loading, g_g, as

$$g_g = g_t + g_L = Kf^2 \tag{20}$$

where, if the g terms are expressed in micromhos and the frequency f in megacycles, the constant K will be in the unit micromhos per megacycle squared. This proportionality factor K is listed in Table 1 for several tube types with the static operating voltages, currents, and signal-grid-to-plate transconductances to which it applies. At frequencies below 25 Mc a conduction term negligible at higher frequencies should be added. This is due to dielectric losses and is proportional to frequency. For single-ended metal tubes this added conduction loss, g_H, is approximately equal to

$$g_H = 0.03f$$

when f is in megacycles and g_H is in micromhos. For the glass miniature and acorn tubes this loss is too small to be measured accurately.

It should be noted that in Table 1 the five entries at the bottom of the list are for mixer or converter use. Those applications where the signal is injected on an outer grid and the local oscillator is injected on the inner grid produce negative loading. In the case of the 6L7, the signal being placed on the grid adjacent to the cathode, positive input loading is produced as in the amplifier applications discussed.

If the grid-plate transconductance is varied from minimum bias to cutoff bias the input loading will vary in proportion to this transconductance and a complicated function of the electron transit times. In Fig. 14 are given in graphical form some measured values of input loading. These data were observed at 100 Mc with the static voltages of Table 1 applied. The grid-plate transconductances were varied by changing the signal-grid bias. To compute the input loading at any frequency, f in megacycles, multiply the loading at 100 Mc by the ratio $(f/100)^2$.

Table 1. Values of K, C_{gk}, and ΔC Input for Several Tube Types

Tube Type	Plate Voltage, volts	Screen Voltage, volts	Signal-grid Bias, volts	Plate Current, ma	Screen Current, ma	Grid-plate Transconductance or Conversion Transconductance, μmhos	C_{gk},* μμf	ΔC Input Change in Input Capacitance,† μμf	Input Loading Constant K, μmhos per Mc²
6SJ7	250	100	−3	3.0	0.8	1650	2.5	1.0	0.053
6SK7	250	100	−3	9.2	2.6	2000	2.1	1.2	0.050
6SH7	250	150	−1	10.8	4.1	4900	4.2	2.3	0.063
6SG7	250	125	−1	11.8	4.4	4700	3.7	2.3	0.060
6AB7	300	200	−3	12.5	3.2	5000	3.6	1.8	0.079
6AC7	300	150	−2	10.0	2.5	9000	6.4	2.4	0.175
9001	250	100	−3	2.0	0.7	1400	1.7	0.5	0.0062
9003	250	100	−3	6.7	2.7	1800	1.4	0.5	0.0066
6AK5	150	120	−2	7.5	2.5	5000	2.6	1.1	0.0134
6AG5	250	150	−1.8	7.0	2.0	5000	3.9	1.4	0.033
6BA6	250	100	−1.2	11.0	4.2	4400	3.5	2.2	0.060
6AU6	250	150	−1	10.8	4.3	5200	3.5	2.5	0.076
954	250	100	−3	2.0	0.7	1400	1.5	0.5	0.005
6J7	250	100	−3	2.0	0.5	1225	2.5	1.0	0.05
6K7	250	100	−3	7.0	1.7	1450	2.1	1.2	0.05
6A8	250	100	−3	3.5	2.7	550			−0.05
6SA7 ‡	250	100	0	3.5	8.5	450			−0.03
6SA7 §	250	100	−2	3.5	8.5	450			−0.03
6K8	250	100	−3	2.5	6.0	350			−0.08
6L7	250	100	−3	2.4	7.1	375			0.15

* This is the capacitance between grid and cathode with all voltages applied and grid biased to plate-current cutoff.

† This is the increase in input capacitance as the grid bias is varied from cutoff to the plate current indicated in this table.

‡ Self-excited.

§ Separately excited.

The reactive component of input admittance is essentially capacitative since the reactances due to the lead inductances to the several tube elements are small up to frequencies of the order of 100 Mc. This input capacitance consists, in the typical grid-controlled amplifiers, of the static paralleling capacitances between the signal grid and all grounded elements plus a variable capacitance which is independent of frequency but changes with tube cathode current and therefore varies with the tube gain.

The variable component of input capacitance may be broken down into two components. One known as the Miller effect is due to the grid-plate or feedback capacitance, C_{gp}. With a pure resistance output load in the plate circuit or one tuned to resonance with the signal-input frequency the added input capacitance is

$$C_{\text{input}} = (1 + A)C_{gp} \tag{21}$$

where A is the voltage gain of the stage. In a triode this component may be of the order of 10 to 100 times as large as the other input capacitance components. Since the grid-plate capacitance of a pentode r-f amplifier tube is of the order of 0.001 that of a triode, this Miller effect can be made negligibly small with a properly designed pentode amplifier.

A second source of input-capacitance variation is due to the space charge in the cathode-grid region. Theory predicts that the capacitance between grid and cathode, C_{gk}, should increase by 33 $^1/_3$ per cent when space-charge-limited current flows as compared to no current flow. Measurements show that there is an increase with current flow but not an abrupt one, the difference being due in all probability to the theory's assumptions of simple geometry and zero initial velocity of emitted electrons not being fulfilled. In Fig. 15 are shown graphically the results of measurements of input-capacitance increases

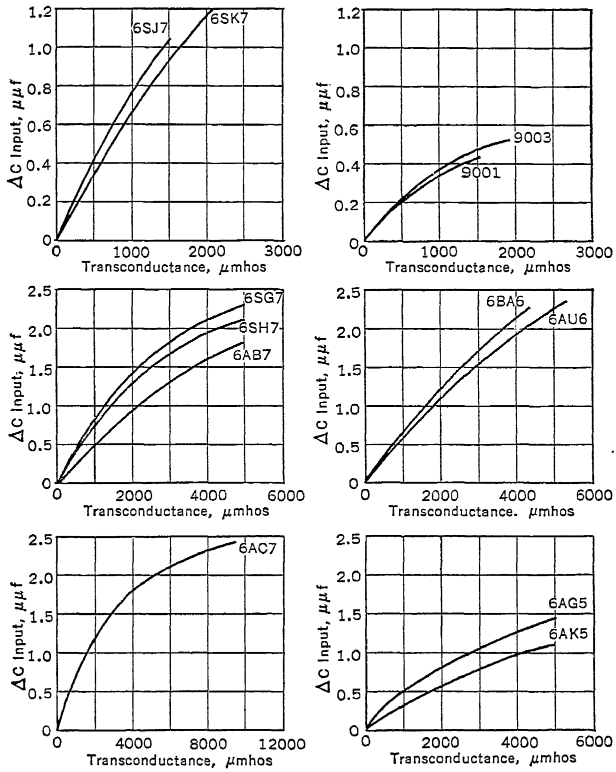


FIG. 15. Increase in Input Capacitance vs. Grid-plate Transconductance for a Group of R-f Pentode Amplifying Tubes

against grid-plate transconductance for a group of typical r-f amplifier pentodes. These measurements were made with the plate grounded to radio frequencies at 100 Mc. Static voltages are as indicated in the previous chart, where the maximum input-capacitance change has also been entered. The transconductance was varied by changing the grid bias.

COMPENSATING FOR INPUT-ADMITTANCE CHANGES. The undesirable changes of input admittance as the gain of an amplifier tube is varied may be made reasonably constant by the introduction of negative feedback through the use of a small unby-passed cathode resistance, r_k. For the input capacitance to have the same value at full transconductance as it has at cutoff, the value of r_k in ohms is given by

$$r_k = \frac{\Delta C_{\text{Input}}}{C_{gk}} \frac{1}{g_{gk}} \tag{22}$$

when C_{gk} is the capacitance at cutoff between grid and cathode, ΔC_{Input} is the increase in input capacitance from cutoff to maximum transconductance, and g_{gk} in mhos is the grid-to-cathode maximum transconductance if the cathode resistance were by-passed with a large capacitor. Both fixed and maximum variational capacitance values are given in Table 1. Note that this compensation holds for any frequency.

For the input conductance to have the same value at cutoff as it has at full gain, the value of the unbypassed cathode resistor, r_k, in ohms is given by

$$r_k = \frac{10^6}{2g_{gk}}\left[-1 + \sqrt{1 + \frac{Kg_{gk}}{\pi^2 C_{gk}^2}}\right] \tag{23}$$

where g_{gk} in micromhos is grid-cathode transconductance, C_{gk} in micromicrofarads is grid-cathode capacitance, and K in micromhos per megacycle squared is the frequency coefficient of input loading. This compensation is also independent of frequency.

Generally r_k will not have the same value for both capacitance and conductance compensation, but practically they are of the same order of magnitude so that correcting for one will usually improve the other. In Fig. 16 are shown graphically the effects on two tube types of adding several different values of r_k.

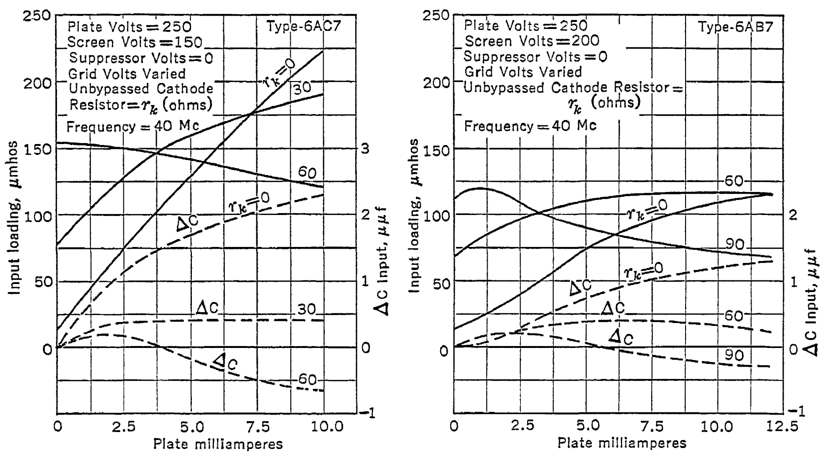


FIG. 16. Input-capacitance Change and Input Loading vs. Plate Current for the 6AB7 and 6AC7 Tubes. (Courtesy *RCA Review*.)

It should be noted that the unbypassed cathode resistance, because it produces degenerative amplification, reduces the gain to $1/(1 + g_{gk}r_k)$ of its value when r_k is by-passed.

NOISE GENERATED IN VACUUM TUBES. The vacuum tube is a noise generator having several possible sources. One important type, known as the "shot effect," stems from the fact that the electron current consists of discrete particles which leave the cathode in a random fashion, producing fluctuation currents uniformly distributed over all frequencies. The "flicker effect" is a low-frequency phenomenon caused by small emitting areas of the cathode constantly changing their emission characteristics. This effect is small compared to the shot effect. In tubes having more than one collector element, such as the screen and plate of a pentode, the random division of current produces uniform noise currents over the whole frequency spectrum of a tube's output or plate current. Other sources are positive-ion-emission currents, positive-ion currents produced as the result of gas ionization, and secondary-electron emission. In the low-frequency region may also be found microphonics due to the motion of the tube elements, and hum resulting from the use of an a-c power source for heating the cathode. Associated with the input loading and therefore present only at high frequencies is another noise source which may add appreciable noise above, say, 30 Mc.

The thermal agitation or shot effect for a temperature-limited current, I, of a diode produces a mean-square fluctuation current $\overline{i^2}$, measured in a frequency band width Δf, as given by the equation

$$\overline{i^2} = 3.18 \times 10^{-19} I \,\Delta f \tag{24}$$

where i and I are expressed in amperes and Δf is in cycles per second. It is often more convenient to express the tube-noise generators in terms of an equivalent resistance which at room temperature produces the same noise as the hot-cathode vacuum tube. The mean-square thermal agitated fluctuation current for any short-circuited pure resistance is

$$\overline{i^2} = \frac{4KT\,\Delta f}{r} \tag{25}$$

where r is the ohmic value of the resistance; K is Boltzmann's constant, 1.372×10^{-23} joule per degree; T is the temperature of the resistance in degrees Kelvin; and i is in amperes. Similarly the mean-square fluctuation voltage, $\overline{e^2}$, across the open-circuit terminals of a pure resistance is

$$\overline{e^2} = 4KTr\,\Delta f$$

where e is now given in volts. If room temperature is assumed to be 290 deg K (81 deg Fahr) the above expressions become

$$\overline{i^2} = \frac{1.59 \times 10^{-20}\,\Delta f}{r} \tag{26}$$

$$\overline{e^2} = 1.59 \times 10^{-20} r\,\Delta f \tag{27}$$

By direct substitution the equivalent-noise resistance of the diode with temperature-limited current flowing is

$$r_{\text{eq}} = \frac{1}{20I} \tag{28}$$

Such a diode is conveniently used as a noise-signal source generator.

With the diode current limited by space charge, and provided that I is small compared to the total available emission current, the fluctuation noise currents are reduced so that the equivalent noise resistance for the diode becomes

$$r_{\text{eq}} = \frac{1}{30I} \tag{29}$$

In triodes and pentodes, shot-effect noise is present as in the diodes and may be represented by a noise-equivalent resistance whose thermal-agitation noise at room temperature is equal to the tube noise referred to the control grid of the tube. In pentodes the random current distributions between the screen grid and anode produce noise usually several times greater than the thermal noise. In the following expressions for equivalent-noise resistances, for the pentode, the first term in the parenthesis is due to shot effect and the second term is due to screen current fluctuation:

For triode amplifiers,

$$r_{\text{eq}} = \frac{2.5}{g_m} \tag{30}$$

For pentode amplifiers,

$$r_{\text{eq}} = \frac{I_b}{I_b + I_{c2}}\left(\frac{2.5}{g_m} + \frac{20I_{c2}}{g_m^2}\right) \tag{31}$$

For triode mixers,

$$r_{\text{eq}} = \frac{4}{g_c} \tag{32}$$

For pentode mixers,

$$r_{\text{eq}} = \frac{I_b}{I_b + I_{c2}}\left(\frac{4}{g_c} + \frac{20I_{c2}}{g_c^2}\right) \tag{33}$$

For multigrid converters (with inner-grid or outer-grid injection),

$$r_{\text{eq}} = 20I_b\,\frac{(I_0 - I_b)}{I_0 g_c^2} \tag{34}$$

The following approximate relationships for triode and pentode mixers, when both the oscillator and signal frequencies are injected in the grid adjacent to the cathode, are useful when the data required for eqs. (32) and (33) are not available. The values "as amplifier" refer to conditions at the peak of the assumed oscillator cycle which is usually very close to zero grid-bias.

g_c (as converter) $= g_m/4$ (as amplifier)
I_b (as converter) $= I_b/4$ (as amplifier)
I_{c2} (as converter) $= I_{c2}/4$ (as amplifier)
r_{eq} (as pentode converter) $= 4r_{eq}$ (as amplifier)
r_{eq} (as triode converter) $= 6.5r_{eq}$ (as amplifier)

Conversion from noise-equivalent resistance to noise-equivalent rms voltage is effected by use of eq. (27):

$$\sqrt{\overline{e^2}} = 1.3 \times 10^{-10}\sqrt{r_{eq}\,\Delta f} \tag{27a}$$

Conversion from noise-equivalent resistance to noise-equivalent rms current is effected by use of eq. (26):

$$\sqrt{\overline{i^2}} = 1.3 \times 10^{-10}\sqrt{\frac{\Delta f}{r_{eq}}} \tag{26a}$$

The symbols in the above equations have conventional significance whose definitions are:

r_{eq}, noise-equivalent resistance, ohms.
g_m, grid-plate transconductance, mhos.
g_c, conversion transconductance (frequency converters and mixers), mhos.
I_b, average plate current, amperes.
I_{c2}, average screen-grid current, amperes.
I_0, average cathode current, amperes.
$\sqrt{\overline{e^2}}$, noise-equivalent rms voltage for band width Δf, volts.
$\sqrt{\overline{i^2}}$, noise-equivalent rms current for band width Δf, amperes.
Δf, effective band width, cycles.

In Table 2 is a listing of representative receiving-tube types showing equivalent-noise values. Several, covering a large range of equivalent-noise resistance values, show measured values all in good agreement with values computed from the above relationships.

Positive-ion noise produced by collision ionization results from residual gas in a vacuum tube. This very undesirable fluctuation noise may be investigated in order to determine the magnitude of grid current (positive-ion current) which may be tolerated from analysis of the following equation for the equivalent-noise resistance, $r_{\text{eq (gas)}}$:

$$r_{\text{eq (gas)}} = \left(20r^2 + A\,\frac{I_0}{g_m{}^3}\right) I_{ig} \quad \text{ohms} \tag{35}$$

where r = grid-circuit resonance impedance, ohms.
I_0 = cathode current, amperes.
I_{ig} = positive-ion current to grid, amperes.
g_m = grid-plate transconductance, mhos.
A = coefficient of the order of 40,000.

The term $20r^2I_{ig}$ represents shot-effect voltage fluctuations produced by the gas current flowing in the grid circuit. It may be increased several fold by induction effects associated with the ion transit time, even at frequencies of a few megacycles. The second term is the noise generator term within the tube. If, for an example, we assume r as 10^5 ohms, g_m as 2000×10^{-6} mho, I_0 as 10^{-2} amp, and a grid current of 10^{-6} amp, the resulting $r_{\text{eq (gas)}}$ is 250,000 ohms. To reduce this equivalent-noise resistance to a value negligible compared to the noise produced by the grid circuit resistance, here assumed to be 100,000 ohms, calls for a reduction of the gas current some 100 fold or to 10^{-2} microampere. Equation (35) is applicable to triodes or pentodes.

There is a source of current fluctuations associated with the component of tube input-conductance produced by electron transit time effects, a source which becomes important when the input frequency is high enough to make the transit-time input conductance relatively large. The random variations in space current will induce current fluctuations in the control-grid circuit, giving rise to grid-voltage fluctuations proportional to the total input impedance (tube and circuit). These induced mean-square grid-current fluctuations may be expressed thus:

$$\overline{i_g{}^2} \approx \frac{20}{3}\left(1 - \frac{\pi}{4}\right) 4KT_kg_g\,\Delta f = 1.43(4KT_kg_g\,\Delta f)$$

Inserting the value for Boltzmann's constant K, and assuming the cathode temperature, T_k, to be 1000 deg K, the above expression simplifies to

$$\overline{i_g{}^2} = 7.5 \times 10^{-20} g_g\,\Delta f \quad \text{amperes} \tag{36}$$

The noise-equivalent rms voltage for the band width Δf appearing at the tube control-grid may be deduced from the above expression; it is

$$\sqrt{\overline{e_g{}^2}} = 2.75 \times 10^{-10}\,\frac{\sqrt{g_g\,\Delta f}}{g + g_g} \quad \text{volts} \tag{37}$$

when g_g in mhos is that portion of the tube input conductance traceable to electronic loading alone; g in mhos is the grid-circuit resonance conductance; and Δf in cycles is the effective band width of the amplifier. As an example, assume the input circuit resonance impedance to be 20,000 ohms so that its reciprocal g is 50×10^{-6} mho; let g_g be 200×10^{-6}

Table 2. Tube-noise Values

(Courtesy *RCA Review*)

Type	Application	Voltages			Currents			Transconductance, micromhos	Noise-equivalent Resistance		Noise-equivalent Input Voltage * microvolts
		Plate Volts	Screen Volts	Bias Volts	Plate milliamperes	Screen milliamperes	Cathode milliamperes		Calculated Ohms	Measured Ohms	
6SK7	Pentode amplifier	250	100	−3	9.2	2.4	11.6	2,000	10,500	9,400–11,500	0.94
6SJ7	Pentode amplifier	250	100	−3	3	0.8	3.8	1,650	5,800	5,800	0.70
6SG7	Pentode amplifier	250	125	−1	11.8	4.4	16.2	4,700	3,300		0.53
6AC7/1852	Pentode amplifier	300	150	−2	10	2.5	12.5	9,000	720	600–760	0.25
956	Pentode amplifier	250	100	−3	5.5	1.8	7.3	1,800	9,400		0.90
1T4	Pentode amplifier	90	45	0	2.0	0.65	2.65	750	20,000		1.3
6SA7	Frequency converter	250	100	0	3.4	8.0	11.9	450 §	240,000	210,000	4.5
6K8	Frequency converter	250	100	−3	2.5	6.0	8.5 ‡	350 §	290,000		4.9
1R5	Frequency converter	90	45	0	0.8	1.8	2.75	250 §	170,000		3.8
6L7	Pentagrid mixer	250	100	−3	2.4	7.1	9.5	375 §	255,000	210,000	4.6
6J5	Triode amplifier	250	....	−8	9.0	....	....	2,600	960	1,250	0.28
955	Triode amplifier	180	....	−5	4.5	....	....	2,000	1,250		0.32
6AC7/1852	Triode amplifier	150	150	−2	....	....	12.5	11,200	220	200	0.14
6AC7/1852	Pentode mixer	300	150	−1 †	5.2	1.3	6.5	3,400 §	2,750	3,000	0.48
6SG7	Pentode mixer	250	125	−1 †	3.0	1.1	4.1	1,180 §	13,000		1.0
956	Pentode mixer	250	100	−1 †	2.3	0.8	3.1	650 §	33,000		1.7
6J5	Triode mixer	100	....	−1 †	2.1	....	....	620 §	6,500		0.74
6AC7/1852	Triode mixer	150	150	−1 †	....	....	6.5	4,200 §	950		0.28
955	Triode mixer	150	....	−1 †	2.8	....	....	660 §	6,100		0.72

* For effective band width of 5,000 cycles.
† At peak of oscillator cycle.
‡ Hexode section only. Triode section takes its current from a separate part of the cathode.
§ Conversion transconductance value.

mho and assume the effective band width to be 5000 cycles. The rms noise voltage appearing at the signal grid is then 1.1×10^{-6} volt, certainly not negligible for a high-gain amplifier!

DISTORTION INTRODUCED BY R-f AMPLIFIER TUBES. Owing to the inherent curvature of the plate-current vs. signal-grid-voltage characteristic of all r-f amplifier tubes, there are present in the tube output three types of distortion when the input signal has the form of an amplitude-modulated carrier. By means of a Taylor expansion series (see also Section 5, articles 16–24), the plate current about any given operating point may be written

$$i_b = I_{b0} + e_g \frac{\partial i_b}{\partial e_c} + \frac{e_g^2}{2!}\frac{\partial^2 i_b}{\partial e_c^2} + \frac{e_g^3}{3!}\frac{\partial^3 i_b}{\partial e_c^3} + \cdots \tag{38}$$

In this expression $\partial i_b/\partial e_c$ is the slope or transconductance, g_{m0}, of the plate-voltage vs. grid-bias curve, and $\partial^2 i_b/\partial e_c^2$ and $\partial^3 i_b/\partial^3 e_c$ are the second and third derivatives respectively of this same curve. The quantity e_g is the alternating part of the grid voltage or the signal input voltage. If an amplitude-modulated carrier signal voltage of the form

$$e_g = E_1(1 + m_1 \sin pt) \sin \omega_1 t$$

is inserted in the above series, we may neglect all the terms having harmonic frequencies of ω_1: if it is assumed that there is a tuned circuit in the output of the tube which band-passes only those frequencies and the sideband frequencies associated with the carrier frequency ω_1. Two types of distortion become apparent from this operation. The modulation factor, m_1, is changed, and the modulation-frequency component has amplitude distortion which is indicated by the presence of harmonic terms of p, the modulation frequency.

The modulation factor is thereby changed by the ratio

$$\frac{\Delta m}{m_1} = \left[\frac{1}{4}\left(1 - \frac{3}{8} m_1^2\right) \frac{\partial^3 i_b/\partial e_c^3}{g_{m0}}\right] E_1^2 \tag{39}$$

The amplitudes of the second and third harmonics producing amplitude distortion of the modulating frequency are expressible as: (*a*) Second harmonic distortion (ratio of the amplitude of sin $2pt$ to the amplitude of sin pt).

$$\text{Ratio of 2nd harmonic to fundamental} = \left[\frac{3}{16} m_1 \frac{\partial^3 i_b/\partial e_c^3}{g_{m0}}\right] E_1^2 \tag{40}$$

(*b*) Third harmonic distortion (ratio of the amplitude of sin $3pt$ to the amplitude of sin pt).

$$\text{Ratio of 3rd harmonic to fundamental} = \left[\frac{1}{32} m_1^2 \frac{\partial^3 i_b/\partial e_c^3}{g_{m0}}\right] E_1^2 \tag{41}$$

A third type of distortion occurs when a second, and usually an unwanted, signal modulates the desired signal. This is known as "cross modulation."

If there is substituted in the above series, for e_g, $e_g = E_1(1 + m_1 \sin pt) \sin \omega_1 t + E_2(1 + m_2 \sin qt) \sin \omega_2 t$ where the subscripts 1 apply to the desired signal and the subscripts 2 apply to the undesired signal, then the cross-modulation ratio is

$$\text{Cross-modulation ratio} = \left[\frac{1}{2}\frac{m_2}{m_1}\frac{\partial^3 i_b/\partial e_c^3}{g_{m0}}\right] E_2^2 \tag{42}$$

Inspection of the three different distortion types indicates that all are proportional to the square of the signal voltage amplitude and that all also are proportional to the third derivative of plate current to grid voltage and inversely proportional to the transconductance. It should also be noted that, since this analysis discards all higher derivatives than the third, reasonably small signal voltages are implied.

It is rather tedious to obtain sufficient points for an i_b vs. e_1 curve to determine the third derivatives accurately by graphical means. A more practical method is to use a conventional g_m bridge having a calibrated and variable signal grid voltage. The procedure is to apply a small signal, of the order of 0.01 to 0.1 volt, to the signal grid, with which a balance is established which gives a transconductance reading equal to g_{m0}. The bridge is then set off-balance by a predetermined small amount, say 1 to 5 per cent, and the input signal is increased until the bridge is again in balance. The required change in g_{m0} balance may be either positive or negative, depending on the curve shape. It can be shown by application of the Taylor's series that

$$\frac{\Delta g_m}{g_{m0}} \approx \left[\frac{1}{8}\frac{\partial^3 i_b/\partial e_c^3}{g_{m0}}\right] E_0^2 \tag{43}$$

where E_0 is the peak value of the large signal applied to the grid. Substitution of the above into the previous distortion equations results in the following expressions:

$$\text{Per cent change in modulation} = 2\left(1 - \frac{3}{8}m_1^2\right)\frac{\Delta g_m}{g_{m0}} \times 100$$

$$\text{Per cent 2nd harmonic distortion} = \frac{3}{2}m_1\frac{\Delta g_m}{g_{m0}} \times 100$$

$$\text{Per cent 3rd harmonic distortion} = \frac{1}{4}m_1^2\frac{\Delta g_m}{g_{m0}} \times 100$$

$$\text{Per cent cross modulation} = 4\frac{m_2}{m_1}\frac{\Delta g_m}{g_{m0}} \times 100$$

If it is assumed that $m_1 = 1$, that $m_2/m_1 = 1$, and that the transconductance is changed 1 per cent, the above may be expressed to show the relative magnitudes of the various distortions, giving

Modulation change = 1.25%
2nd harmonic distortion = 1.5%
3rd harmonic distortion = 0.25%
Cross modulation = 4%

In Fig. 17 are plotted the signal voltage necessary to change the g_m by 3 per cent vs. the

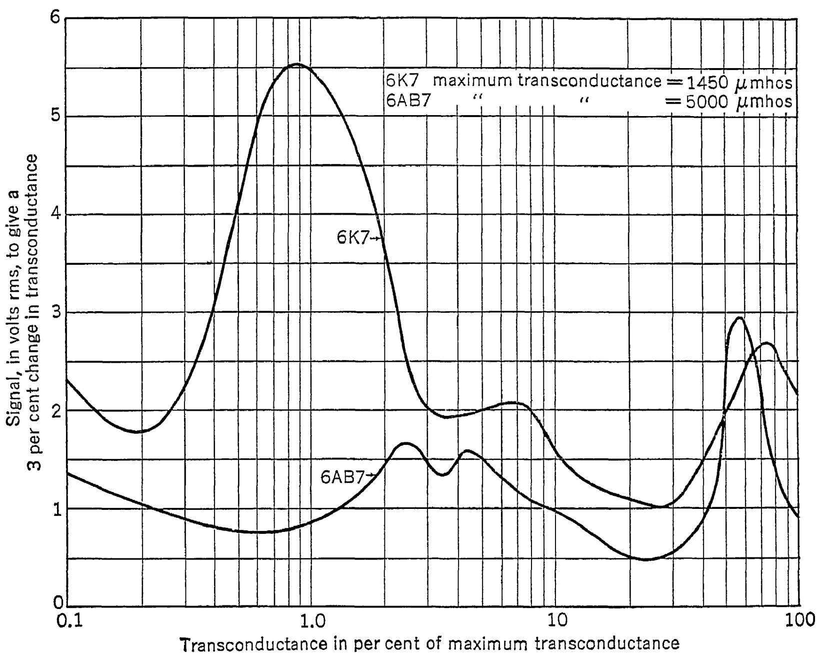


FIG. 17. Comparison of Signal Voltages Applied to the Remote and Semi-remote Cutoff Tube Types 6K7 and 6AB7 Respectively to Produce the Same Amount of Distortion

transconductance in terms of the maximum transconductance. For modulation factors of unity, this corresponds to saying that the signal voltages indicated will produce

Modulation change = 3.75%
2nd harmonic distortion = 4.5%
3rd harmonic distortion = 0.75%
Cross modulation = 12%

The one curve is for a "remote cutoff" tube, 6K7, which requires approximately 45 volts bias to reduce its maximum transconductance of 1500 μmhos to about 10^{-3} this value.

The second curve is for a "semi-remote cutoff" tube, 6AB7, which requires about 22 volts bias to reduce its maximum transconductance from 5000 to 10^{-3} this value. It is apparent that the 6K7 can handle a much larger signal than the 6AB7. The 6AB7 tube is limited to lower signal values because of its higher transconductance value, and if it were designed to have a similar cutoff as the 6K7 it would draw exceedingly large plate currents when the bias was set low for maximum gain. There is thus a practical limitation governing the maximum transconductance and maximum signal-handling capacity that can be designed into a tube.

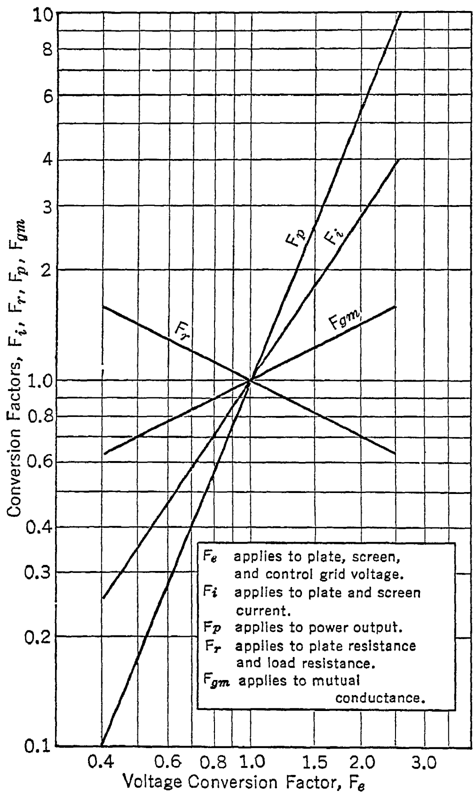


FIG. 18. Conversion Factor for Triode and Pentode

CHANGING OPERATING CONDITIONS. Operating Voltages. It is sometimes necessary to change the operating potentials of an amplifier or a power output triode or pentode tube from the published typical operating conditions. By means of the conversion factor chart, Fig. 18, it is possible to determine the new voltages, currents, transconductance, plate and load resistances, and power output. The curves are based on the fact that, if all applied voltages (except heater voltage) are changed by a factor n, the resulting currents will all be changed by $n^{3/2}$, the g_m by $n^{1/2}$, and the plate resistance by $n^{5/2}$. The accuracy of the chart is reasonably good for small changes, but for large voltage changes exceeding 2.5 to 1 the chart is unsuitable. As an example, assume a pentode with the following typical ratings:

Plate voltage	250 volts
Screen voltage	250 volts
Grid bias	−18 volts
Plate current	32 ma
Screen current	5.5 ma
Plate resistance	70,000 ohms
Transconductance	2,300 μmhos
Load resistance	7,600 ohms
Power output	3.4 watts

It is desired to determine the operation characteristics for a plate voltage of 100 volts. All voltages will have to be changed in the ratio of 100 to 250 (e.g., F_e, the voltage conversion factor, is 0.4). From the chart the new to the old ratios are picked off so that F_i is 0.25, F_p is 0.1, F_{gm} is 0.63, and F_r is 1.6. The new conditions will then be

Plate voltage		100 volts
Screen voltage	250 × 0.4	= 100 volts
Grid bias	−18 × 0.4	= −7.2 volts
Plate current	32 × 0.25	= 8 ma
Screen current	5.5 × 0.25	= 1.4 ma
Plate resistance	70,000 × 1.6	= 112,000 ohm
Transconductance	2,300 × 0.63	= 1,450 μmhos
Load resistance	7,600 × 1.6	= 12,000 ohms
Power output	3.4 × 0.1	= 0.34 watt

It should be noted that this chart cannot be used if only one voltage is varied.

Changes in Heater or Filament Voltage. Changing the heater or filament voltage of a tube will increase or decrease the temperature, current, and power input of the emitting cathode. The curves of Fig. 19 show these relationships, which were based on taking average values of radiation coefficients and resistivity of tungsten, molybdenum, tantalum, and nickel covering temperature ranges of 1000 to 2800 deg. K. With a maximum filament voltage variation of ±25 per cent, engineering accuracy holds for vacuum and gas-filled tubes, the greatest deviation between observed and predicted currents and power input being only 4 per cent.

As an example assume that a 6.3 volt 0.3 amp heater is running with a cathode temperature of 1050 deg K and a heater temperature of 1400 deg K and that the heater voltage is decreased to 5.5 volts or 87.3 per cent of nominal. From Fig. 19 we find that the resulting power input is 80 per cent of 1.89 watts, or 1.51 watts. The new filament current is reduced to 92 per cent of 0.3 amp, or 0.276 amp, and the cathode and heater temperature are reduced to 95.8 per cent, or 1005 and 1340 deg K, respectively.

Maximum Allowable Grid Resistance. Common practice is to apply the d-c bias to the negative control grid of an amplifier through a series grid resistor. It is desirable to make this resistor as high valued as possible since it shunts the signal source and therefore absorbs power and may have detrimental results on the frequency response and loading of the signal source. Its limiting value is usually of the order of 0.1 to 10 megohms and is established by the radio-tube manufacturer on the basis of life tests and maximum expected grid current. Grid current flowing through the grid resistor decreases the grid bias by the IR drop through the resistor and may cause the tube to "run away" since the resulting increase in plate current will produce excessive plate dissipation. By the addition of a cathode self-bias resistor, or the use of a series dropping resistor in the screen lead of a pentode, or the use of a d-c load resistor in the plate lead of a triode, there results some d-c degeneration, thereby making it possible to increase the value of the grid resistor above that indicated as a maximum. Also, if the tube is operated at a reduced value of transconductance, the maximum allowable grid resistance value may be increased.

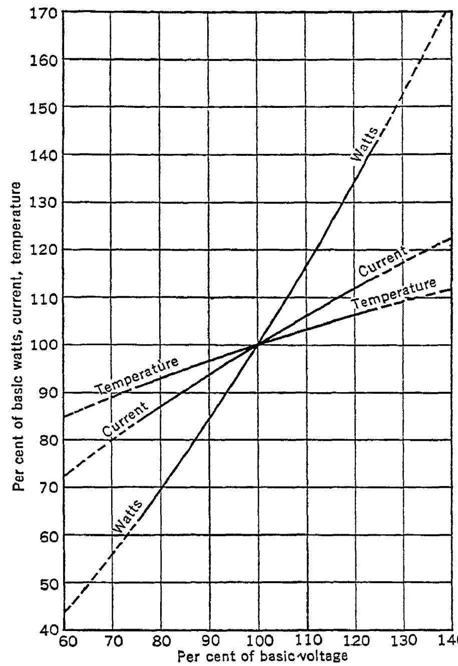


FIG. 19. Chart Giving Wattage, Current, and Temperature of a Filament, Heater, or Cathode at Operating Voltages up to 25 Per Cent above or below Basic Voltage, with Sufficient Accuracy for Most Engineering Purposes. Accuracy drops in dotted regions.

From the following equation it is possible to determine the maximum allowable value of grid resistance, r_{g1}, for any new set of operating conditions differing from those published.

$$r_{g_1} = \frac{\Delta I_k}{\Delta I_{g_1}}\left[\frac{1}{g_k}\left(1 + \frac{r}{r_p + r}\right) + r_k\left(1 + \frac{1}{\mu}\right)\right] \tag{44}$$

where, for a triode,

$\Delta I_k/\Delta I_{g_1}$ is the ratio of plate current change per unit change in grid current;
g_k is the grid-plate transconductance in mhos;
r is the d-c plate load resistance;
r_p is the internal tube plate resistance;
r_k is the cathode self-bias resistance in ohms;
μ is the triode mu;

and where, for a pentode,

$\Delta I_k/\Delta I_{g_1}$ is the ratio of cathode current change per unit change in grid current;
g_k is the signal-grid-to-cathode transconductance in mhos (this may be determined by multiplying the signal-grid-to-plate transconductance by the factor $(I_p + I_{c_2})/I_p$);
r is the screen dropping resistance;
r_p is the internal tube screen resistance;
r_k is the cathode self-bias resistance, in ohms;
μ is the triode connected amplification factor, i.e., the mu from control grid to screen grid of the pentode.

Cathode Volts			1.4	2.0	2.5–5.0	6.3	12.6–117
Kinescopes							
Projection	magnetic deflection					5TP4	
Directly Viewed	magnetic deflection				9AP4 12AP4	7DP4 10BP4	
	electrostatic deflection					7JP4	
Rectifiers (*for rectifiers with amplifier units, see Power amplifiers*)							
Half-wave	vacuum		1B3-GT/ 8016 *			1-v 81 †	12Z3 [35W4 35Y4 35Z4-GT 35Z3 35Z5-GT] 45Z3 45Z5-GT 117Z3
Full-wave	vacuum				5T4, 5W4 [5U4-G, 5X4-G 5Z3] [5Y3-GT 5Y4-G 80] 5Z4 [5V4-G, 83-v]	[6X4, 6X5 6X5-GT, 84/6Z4] 6ZY5-G 7Y4 7Z4	
	mercury-vapor				82 83		
	gas		Cold-Cathode Types: 0Z4, 0Z4-G				
Doubler	vacuum						[25Z5 25Z6 25Z6-GT] 50Y6-GT 117Z6-GT
Diode detectors (*for diode detectors with amplifier units, see Voltage amplifiers and also Power amplifiers*)							
One diode			1A3				
Two diodes						6AL5 [6H6, 6H6-GT] 7A6	12H6 12AL5
Power amplifiers with and without rectifiers, diode detectors, and voltage amplifiers							
Triodes	low-mu	single unit		31 49	2A3 45 46 71-A	6B4-G 10 † 6A3 50 †	
	high-mu	single unit				6AC5-GT	
	high-mu	twin unit	1G6-GT	[1J6-G 19]	53	[6A6, 6N7 6N7-GT] 6Z7-G 79	
	direct-coupled arrangement					6B5	
Beam Tubes	single unit		[1Q5-GT 3Q5-GT ‡] 1T5-GT 3LF4 ‡			6BG6-G [6AQ5 6V6 6V6-GT] [6L6 6L6-G] 6Y6-G 7A5 7C5	14A5 [25L6 35A5 25L6-GT] [35B5 35L6-GT] 50A5 [50B5, 50L6-GT]
	with rectifier						32L7-GT 70L7-GT [117L7/M7-GT 117P7-GT] 117N7-GT
Pentodes	single unit		1A5-GT 1C5-GT 1LA4, 1LB4 [1S4, 3S4 ‡] [3Q4‡, 3V4‡]	[1F4 1F5-G] 1G5-G 1J5-G 33	2A5 47 59	6A4/LA [6AK6, 6G6-G] 6AG7 [6F6, 6F6-G, 6F6-GT, 42] [6K6-GT, 41] 7B5 38 89	[25A6 43]
	with medium-mu triode					6AD7-G	
	with diode and triode		1D8-GT				
	with rectifier						12A7
	twin unit			1E7-G			

* Cathode volts, 1.25. † Cathode volts, 7.5.

FIG. 20. Receiving Tube Classi-

		Cathode Volts	1.4	2.0	2.5–5.0	6.3	12.6–117
Converters & mixers (*for other types used as mixers, see Voltage amplifiers*)							
Converters	pentagrid		1A7-GT 1LA6 1LC6 1R5	[1C6 1C7-G] [1A6 1D7-G]	2A7	[6A7, 6A8 6A8-G, 6A8-GT 6D8-G] [6BE6, 6SA7 6SA7-GT] 6SB7-Y 7B8 7Q7	12A8-GT [12BE6, 12SA7 12SA7-GT] 14B8 14Q7
	triode-hexode					[6K8, 6K8-G]	12K8
	triode-heptode					6J8-G 7J7 7S7	14J7
	octode					7A8	
Mixers	pentagrid					[6L7, 6L7-G]	
Electron-ray tubes							
Single	with remote-cutoff triode					6AB5/6N5 6U5/6G5	
	with sharp-cutoff triode				2E5	6E5	
Twin	without triode					6AF6-G	
Voltage amplifiers with and without Diode Detectors; Triode, tetrode, and pentode detectors; oscillators							
Triodes	medium-mu	single unit	1G4-GT 1LE3, 26 §	[1H4-G 30]	27 56	6C4 [6C5, 6C5-GT] [6P5-GT, 76] [6J5, 6J5-GT] 6L5-G, 7A4, 37	12J5-GT 14A4
		with r-f pentode				6F7	
		with power pentode				6AD7-G	
		with pentode and diode	1D8-GT 3A8-GT				
		with two diodes		[1B5/25S 1H6-G]	55	[6R7, 6R7-GT 6BF6, 6SR7, 6ST7] 7E6 85	[12SR7 12SR7-GT] 14E6
		twin unit				6C8-G [6F8-G, 6SN7-GT] 6J6 7N7 7F8 12AU7	12AH7-GT 12AU7 12SN7-GT 14N7
	high-mu	single unit				[6F5, 6F5-GT 6SF5, 6SF5-GT] 6K5-GT 7B4	[12F5-GT 12SF5]
		with diode	1H5-GT 1LH4				
		with two diodes			2A6	[6SQ7, 6SQ7-GT 6B6-G, 75] 6T7-G, 7B6, 7C6 [6AT6, 6AQ6 6Q7, 6Q7-G 6SZ7, 6Q7-GT]	[12AT6, 12Q7-GT] 14B6 [12SQ7, 12SQ7-GT]
		with three diodes				6S8-GT	
		twin unit				6SC7 6SL7-GT 7F7 12AX7	12SC7 12AX7 12SL7-GT 14F7
Tetrodes	remote cutoff				35		
	sharp cutoff			32	24-A	36	
Pentodes	remote cutoff	single unit	1T4 1P5-GT	34 [1D5-GP 1A4-P]	58	[6K7, 6K7-G 6K7-GT, 78] [6D6 6U7-G] [6BA6, 6SG7] 6BJ6 6AB7/1853 [6S7 6S7-G] 7A7, 7B7 [6SK7 6SK7-GT] 6SS7 7H7 39/44	[12BA6, 12SG7] [12SK7 12SK7-GT] 12K7-GT, 14H7 14A7/12B7
		with triode				6F7	
		with diode				6SF7	12SF7
		with two diodes			2B7	[6B7 6B8, 6B8-G] 7E7 7R7	12C8, 14R7
	sharp cutoff	single unit	1LC5, 1LN5 1L4, 1U4 1N5-GT	[1E5-GP 1B4-P] 15	57	[6J7, 6J7-G, 6J7-GT 6C6, 6W7-G, 77] [6SJ7 6SJ7-GT] [6AU6 6SH7] 6AC7/1852 6AG5 7G7/1232 7C7 7L7 7V7 7W7	[12AU6, 12SH7] 12AW6 [12SJ7 12SJ7-GT] 12J7-GT 14C7
		with diode	1LD5 [1S5, 1U5]				
		with two diodes		[1F6 1F7-G]			

‡ Filament arranged for either 1.4- or 2.8-volt operation. § Cathode volts, 1.5.

fication Chart. (Courtesy RCA.)

All the above factors can usually be obtained from published tube characteristics or are measurable for a given tube type with the exception of the ratio $\Delta I_k/\Delta I_{g_1}$. This ratio indicates that for any given amount of grid current flow there is a definite increase in cathode current flow. If it were possible to manufacture vacuum tubes so that with a negative grid absolutely no grid current flowed, the series-grid resistance could be infinite in value. However, for practical tubes, grid currents of the order of a microampere may exist due to residual gas ionization, thermal and photoelectric emission, and leakage currents. This ratio, $\Delta I_k/\Delta I_{g_1}$, for a given tube may be determined from the published maximum allowable grid resistance by substituting it in the above equation, and once determined it may be used to determine a new value of maximum resistance for a new operating condition.

Example. A pentode with fixed bias and fixed screen voltage has indicated a maximum allowable grid resistance of 0.2 megohm. What is the maximum grid resistance with full self-bias and a series dropping resistance from a 300-volt supply? What is the maximum grid resistance if the bias is increased so that the cathode current is reduced to one-tenth its maximum value?

For $r_{g1} = 0.2$ megohm,

$$\begin{aligned} E_{g1} &= -2\,1/2 \text{ volts} \\ E_{c2} &= 100 \text{ volts} \\ E_{c3} &= 0 \text{ volts} \\ E_b &= 250 \text{ volts} \\ g_{1-p} &= 5000\ \mu\text{mhos} \\ \mu_{12} &= 25 \\ r_2 &= 25{,}000 \text{ ohms (internal screen resistance)} \\ i_{c2} &= 2.5 \text{ ma} \\ i_p &= 10 \text{ ma} \end{aligned}$$

The cathode transconductance, g_{1k}, is determined by,

$$g_k = g_{1-p}\frac{(I_p + I_{c_2})}{I_p} = \frac{5000}{10^{-6}}\frac{(12.5)}{10} = 6250 \times 10^{-6} \text{ mho} \tag{45}$$

Then $r_1 = \dfrac{\Delta I_k}{\Delta I_{g_1}}\left(\dfrac{1}{g_k}\right)$, since there are no self-bias or screen dropping resistors used. Then

$$\frac{\Delta I_k}{\Delta I_{g_1}} = r_1 g_k = (0.2 \times 10^6)(6250 \times 10^{-6}) = 1250 \tag{46}$$

For the new conditions the self-bias and series dropping screen resistors are

$$r_k = 2.5/(12.5 \times 10^{-3} \text{ amp}) = 200 \text{ ohms}$$

$$r_{g_2} = (300 - 100)/(2.5 \times 10^{-3} \text{ amp}) = 160{,}000 \text{ ohms series dropping screen resistor}$$

The new maximum grid resistance is

$$\begin{aligned} r_{g_1} &= 1{,}250\left[\frac{10^6}{6250}\left(1 + \frac{160{,}000}{160{,}000 + 25{,}000}\right) + 200\left(1 + \frac{1}{25}\right)\right] \\ &= 632{,}000 \text{ ohms} \end{aligned} \tag{47}$$

At one-tenth normal cathode current the new bias is computed to be approximately -3.7 volts, assuming that the tube obeys the three-halves power of effective voltage law. The new values are then

$$\begin{aligned} r_k &= 3.7/(1.25 \times 10^{-3} \text{ amp}) = 2950 \text{ ohms} \\ r_{12} &= 160{,}000 \times 10 = 1.6 \text{ megohms series screen resistor} \\ r_{g_2} &= 25{,}000\ (10)^{1/3} = 53{,}000 \text{ internal screen resistance} \\ g_k &= 6250 \times 10^{-6}/10^{1/3} = 2910 \times 10^{-6} \text{ mho} \end{aligned}$$

From the above the new maximum grid resistance is

$$\begin{aligned} r_{g_1} &= 1250\left[\frac{10^6}{2910}\left(1 + \frac{1.6}{1.6 + 0.053}\right) + 2950\left(1 + \frac{1}{25}\right)\right] \\ &= 4.7 \text{ megohms} \end{aligned} \tag{48}$$

RECEIVING TUBE CLASSIFICATION CHART. Figure 20 classifies the commonly used receiving tubes according to their functions and their cathode voltages. It is arranged to permit quick determination by the tube user of the type designations of tubes applicable to specific design requirements. Types having similar characteristics and in the same cathode-voltage groups are bracketed.

6. TYPICAL VACUUM-TUBE CHARACTERISTIC CURVES

The following vacuum-tube characteristic curves were selected as representative of the *triode* (type 10), the *tetrode* (type 865), and the *suppressor-grid pentode* with normal (type 57) and with *remote-cut-off* control grid (type 58). The data were furnished by the RCA Radiotron Division, RCA Mfg. Co., Inc.

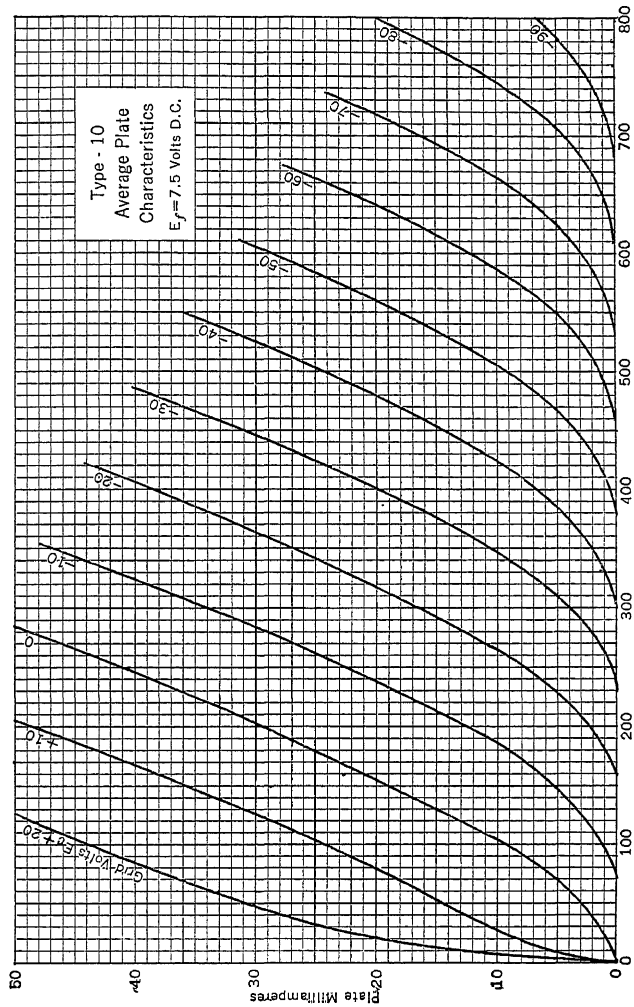


FIG. 21. Typical Triode (Type 10)

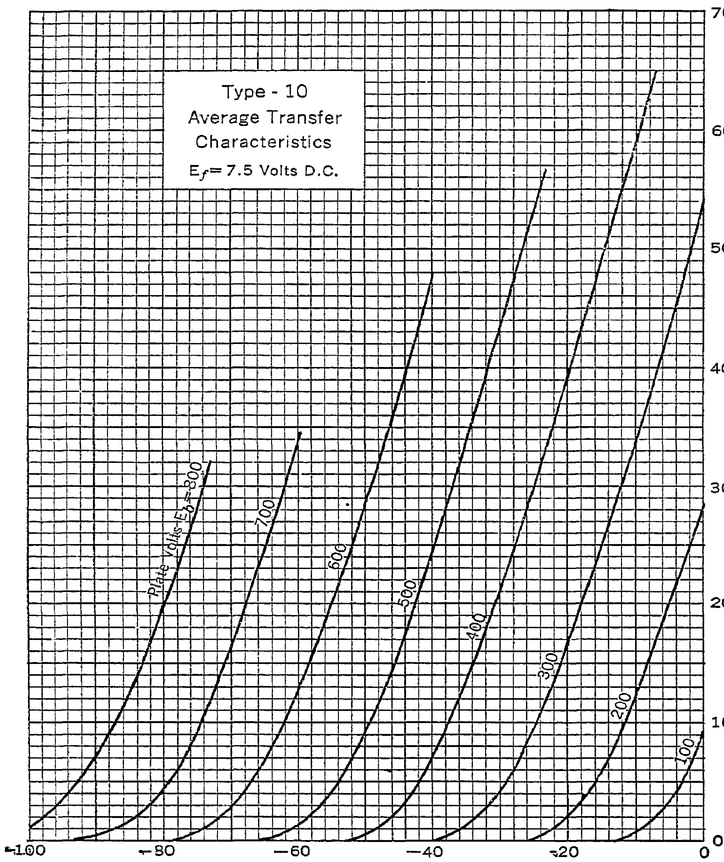


FIG. 22. Typical Triode (Type 10)

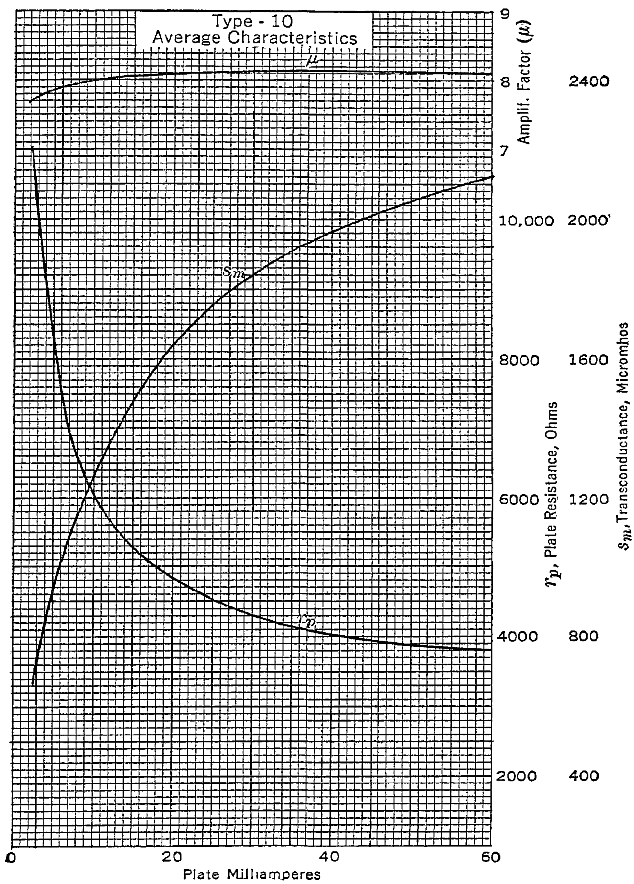


FIG. 23. Typical Triode (Type 10)

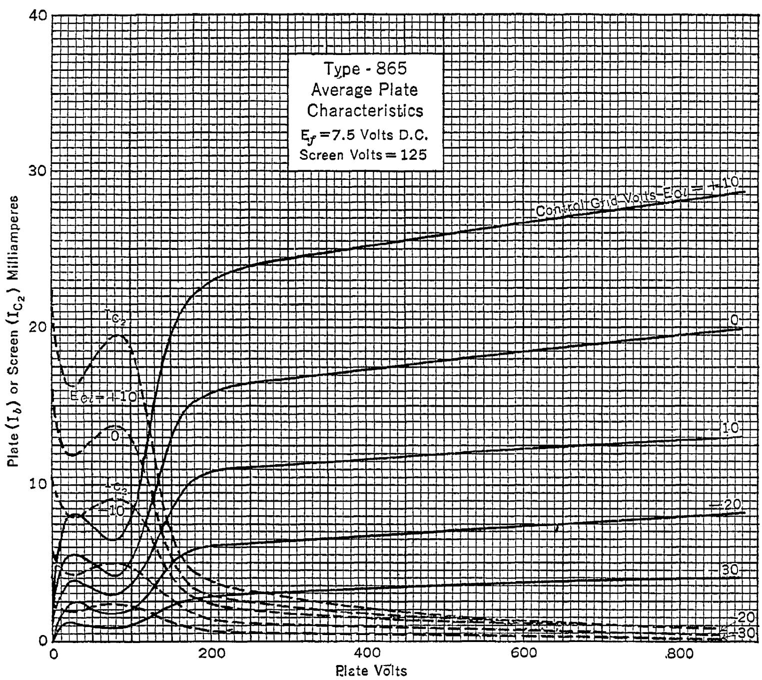


FIG. 24. Typical Tetrode Characteristics (Type 865)

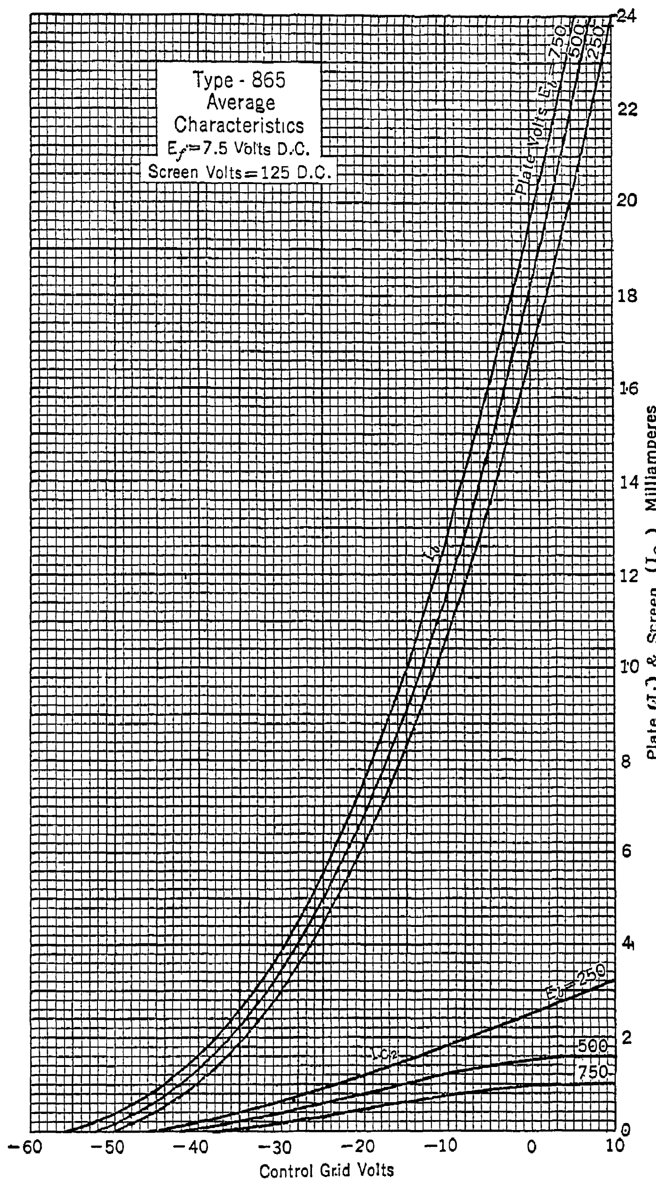


FIG. 25. Typical Tetrode Characteristics (Type 865)

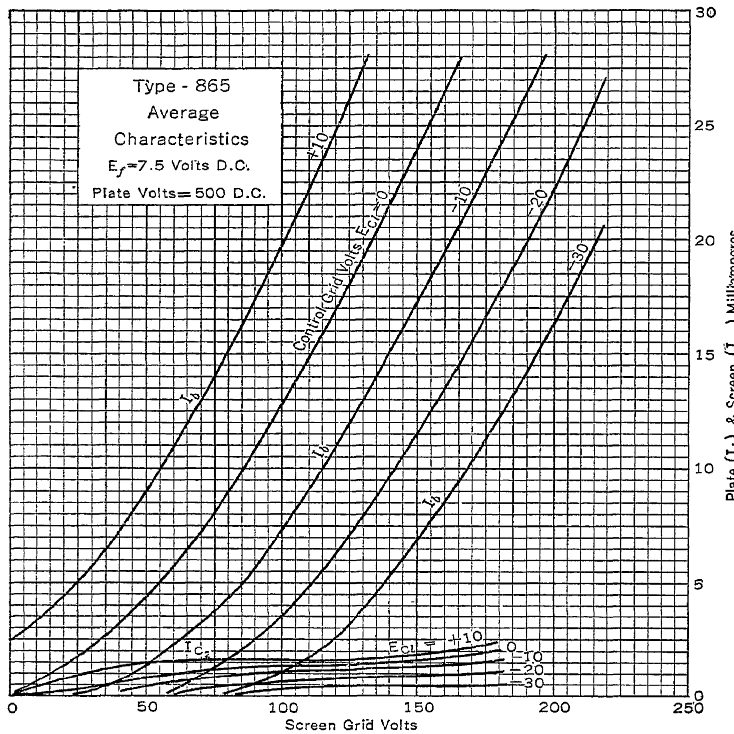


Fig. 26. Typical Tetrode Characteristics (Type 865)

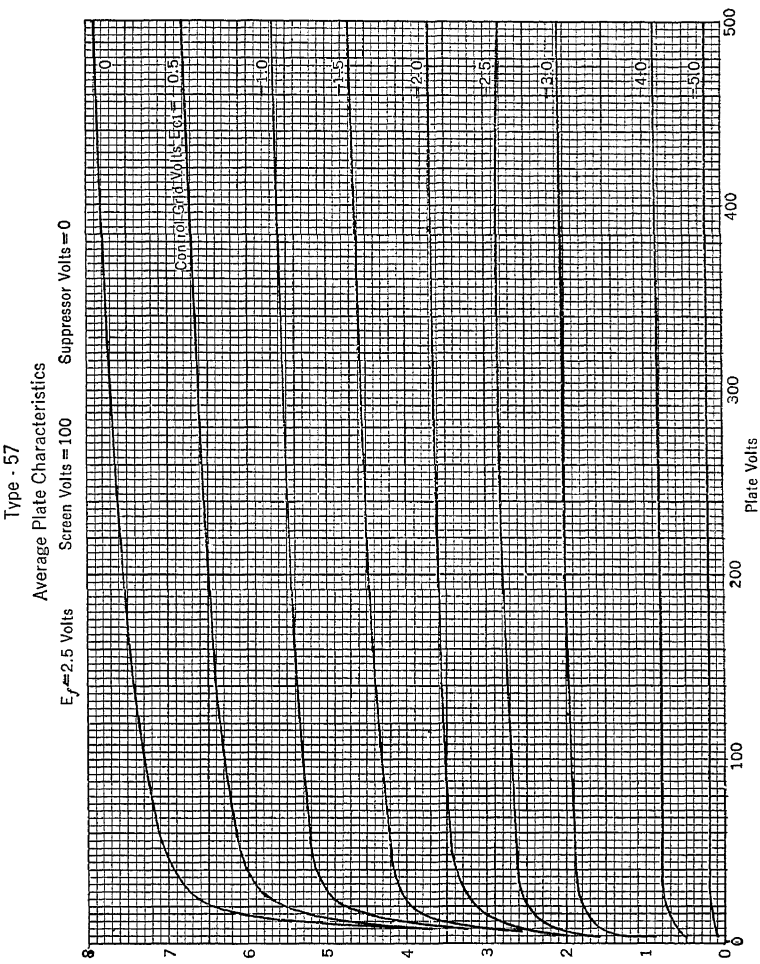


FIG. 27. Typical Pentode Characteristics (Type 57)

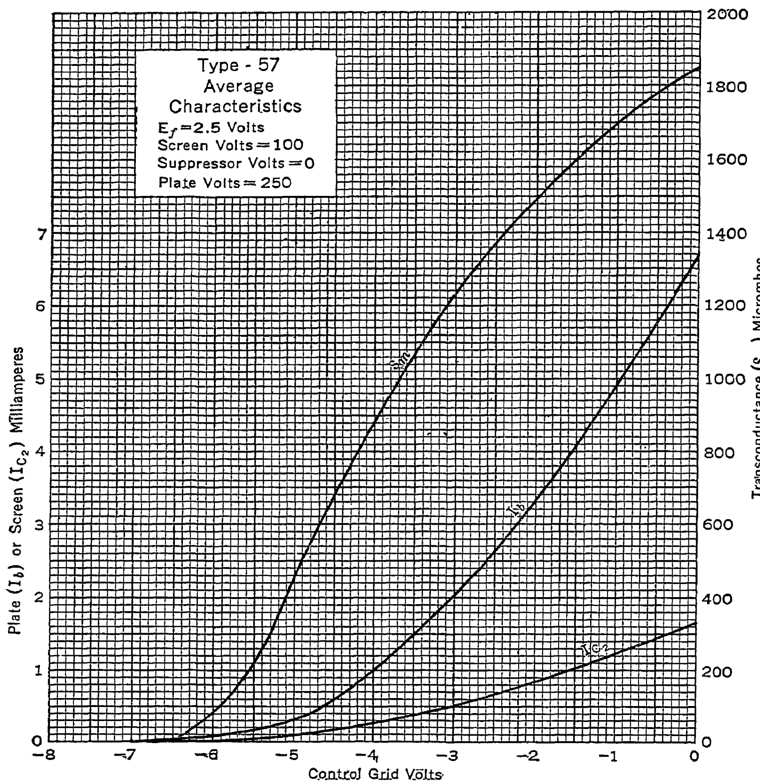


FIG. 28. Typical Pentode Characteristics (Sharp Cutoff Type 57)

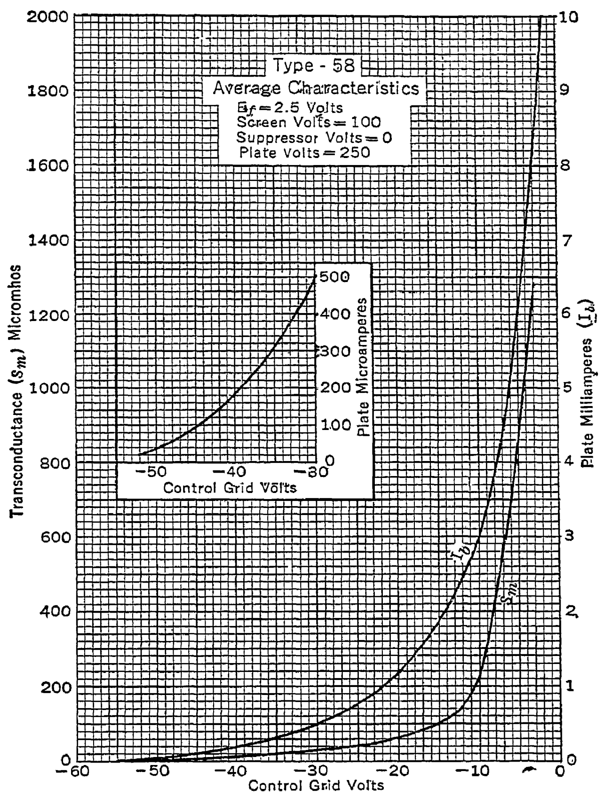


FIG. 29. Typical Pentode Characteristics (Remote Cutoff Type 58)

BIBLIOGRAPHY

Llewellyn, F. B., *Electron-Inertia Effects*, Cambridge University Press, Cambridge (1943).
Ferris, W. R., Input Resistance of Vacuum Tubes as Ultra-high-frequency Amplifiers, *Proc. I.R.E.*, Vol. 24, 82 (January 1936).
North, D. O., Analysis of the Effects of Space Charge on Grid Impedance, *Proc. I.R.E.*, Vol. 24, 108 (January 1936).
Strutt, M. J., and A. Van der Ziel, The Causes for the Increase of Admittances of Modern High-frequency Amplifier Tubes on Short Waves, *Proc. I.R.E.*, Vol. 26, 1011 (August 1938).
Rothe, H., and W. Kleen, *Grundlagen and Kennlinien der Electronenröhren*, Akademische Verlagsgesellschaft, Leipzig (1943). Also by the same authors and publisher, *Elecktronenröhren als Anfangsstufen-Verstarkes* (1940).
Langmuir, I., and K. T. Compton, Electrical Discharges in Gases, Part II. Fundamental Phenomena in Electrical Discharges. *Rev. Modern Phys.*, Vol. 3, 191 (April 1931). This is an excellent paper.
Herold, E. W., The Operation of Frequency Converters and Mixers for Superheterodyne Reception, *Proc. I.R.E.*, Vol. 30, 84 (February 1942).
Thompson, B. J., D. O. North, and W. A. Harris, Fluctuations in Space-charge-limited Currents at Moderately High Frequencies, *R.C.A. Rev.*, Vol. 4, 269–285 (January); Vol. 4, 441–472 (April); Vol. 5, 106–124 (July); Vol. 5, 240–260 (October 1940); Vol. 5, 371–388 (January 1941).
North, D. O., and W. R. Ferris, Fluctuations Induced in Vacuum-tube Grids at High Frequencies, *Proc. I.R.E.*, Vol. 29, 49–50 (February 1941).
Herold, E. W., Superheterodyne Converter System Considerations in Television Receivers, *R.C.A. Rev.*, Vol. 4, 324–337 (January 1940). This article gives general expressions for conversion transconductance and equivalent noise.
Haller, C. E., Filament and Heater Characteristics, *Electronics*, July 1944, pp. 126–130.
Ballantine, S., and H. A. Snow, Reduction of Distortion and Cross-talk in Radio Receivers by Means of Variable-mu Tetrodes, *Proc. I.R.E.*, Vol. 18, 12, pp. 2102–2107 (December 1930).
Herold, E. W., R-f Distortion or Cross-modulation of Pentode Amplifier Tubes. *Electronics*, April 1940, pp. 82–88.

MAGNETRONS

By W. B. Hebenstreit

In this section only those magnetrons of circular cylinder geometry will be considered. A cylindrical anode is coaxial to an emitting cathode, and the elements are mounted in a vacuum envelope. A static magnetic field is parallel, or nearly parallel, to the axis of the cathode, and a d-c potential applied between the cathode and anode sets up a radial, static electric field. Under conditions of oscillation, the electrons also interact with an a-c field.

7. THE NON-OSCILLATING MAGNETRON

The simplest example is the non-oscillating solid anode magnetron. Under the influence of the electric field, the electron is impelled to move toward the anode. The magnetic field results in a force on the electron which is normal both to the direction of motion and to the direction of the magnetic field. Theoretically, there is a minimum critical voltage, V_c, called the cutoff voltage, for each value of the magnetic field, B, for which electrons will just reach the anode. The formula * for the cutoff voltage is

$$V_c = \frac{eB^2 r_a^2}{8m}(1 - \sigma^2)^2 \tag{1}$$

where e is the electronic charge, m is the electronic mass, r_a is the anode radius, and σ is the ratio of the cathode radius to the anode radius. In this formula it is assumed that the electrons have zero velocity at the cathode; in addition the relativistic effects, which become significant at high voltages, are neglected. Since V_c is proportional to B^2, the locus of eq. (1) for any given geometry is often referred to as the cutoff parabola.

8. THE OSCILLATING MAGNETRON

Principal current interest lies in the use of a magnetron as a self-excited oscillator. A convenient classification of oscillating magnetrons can be made by distinguishing among the several ways in which electrons interact with the a-c fields to sustain oscillations. For this purpose, three types of interactions can be identified.

Type I. The Negative Resistance Magnetron. If the anode of the magnetron is split into halves and one half is raised to a higher potential than the other, under certain conditions, most of the electrons will go to the plate of lower potential. In this event, a negative

* Unless otherwise indicated, mks units will be used in articles 7–9.

resistance exists between the two halves of the anode, and oscillations will be sustained in a tank circuit which is connected between them.

Type II. The Cyclotron Frequency Magnetron. In the neighborhood of cutoff (see eq. [1]), the solid anode magnetron will sustain oscillations if the terminals of an L-C tank circuit are connected between the cathode and anode. The wavelength of the oscillations which will be sustained is given by

$$\lambda B = \text{Constant} \tag{2}$$

The most commonly observed value of the constant is about 15,000 centimeter-gausses. The disadvantage of this type of magnetron is that the electronic efficiency is very low.

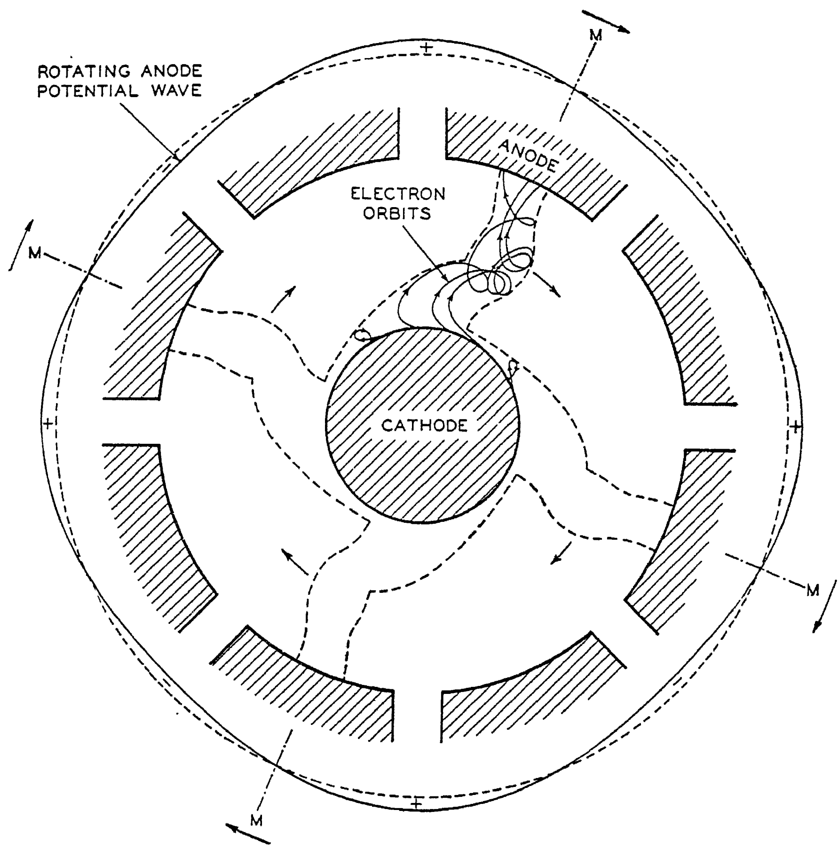


FIG. 1. Approximate Configuration of a Space-charge Cloud of an 8-segment Magnetron Operating in the π-mode

The electrons that initially absorb energy from the field are removed from the interaction space within approximately one cycle either by striking the anode or by being returned to the cathode. Those electrons that initially yield energy to the field stay in the interaction space for a longer time. However, after they have lost their energy they begin to reabsorb it again unless some method is provided for their removal. This is sometimes done either by tilting the magnetic field slightly with respect to the axis of the cathode or by installing electrodes at the ends of the interaction space which are made positive with respect to the cathode. In either method the electrons that have given energy to the r-f field are drawn off at the ends if the interaction space is short enough.

Type III. The Traveling Wave Magnetron. In this magnetron, the anode consists of a number of segments. In operation, an r-f standing wave pattern exists in the interaction space between the cathode and anode. In general, standing wave patterns may be thought of as being composed of two waves traveling in opposite directions. The standing wave pattern in the interaction space is composed of two traveling waves rotating in opposite directions around the interaction space.

Oscillations are sustained by interaction between one of the traveling-wave components and the electronic stream. The electronic stream assumes the shape of a spoked cloud which is centered on the axis of the tube as is indicated schematically in Fig. 1. The spokes wheel around the interaction space in synchronism with one of the components of the rotating wave in such a phase that the spokes are in a retarding tangential * electric field. That is to say that the electrons in the spokes are yielding energy to the r-f field. Those electrons that come from the cathode in such a phase as initially to absorb energy from the field are returned to the cathode after only one orbital loop. Figure 1 shows the computed paths of electrons emitted at several different phases. The paths are drawn as they would appear to an observer stationed in a system of coordinates which is centered at the axis of the tube and which is rotating with the same angular velocity as the rotating wave.

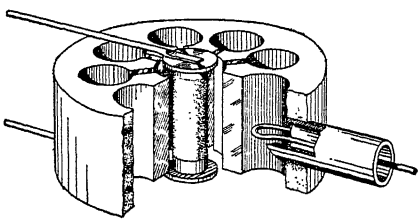


FIG. 2a. Schematic Representation of a Hole and Slot Magnetron

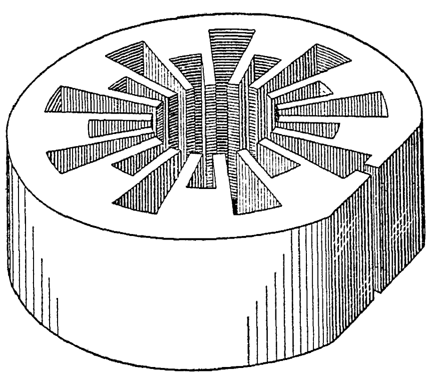


FIG. 2b. Schematic Representation of a Rising Sun Anode Block

This selective mechanism, that is, the mechanism by which the unfavorable electrons are rejected by being returned to the cathode in a relatively short time and the favorable electrons are grouped into spokes which stay in a retarding r-f field, results in very high electronic efficiencies. Electronic efficiencies of 60 per cent are not uncommon.

Associated with the segments of the anode is a system of resonators. The resonator system may assume any one of a variety of forms. One of the commonest forms is illustrated in Fig. 2a. This figure is a schematic representation of the hole-and-slot-type magnetron. Each hole and slot can be thought of as an L-C tank circuit with an associated resonant frequency.

In the rising sun structure, illustrated in Fig. 2b, resonators of one resonant frequency alternate with resonators of another frequency. In normal operation, the operating frequency of the ensemble is roughly midway between the resonant frequencies of the two sets.

From the standpoint of practical application, the traveling-wave magnetron is by far the most important type of magnetron oscillator. It will be the exclusive concern of article 9, and, unless explicitly stated, the term "magnetron" will mean a traveling-wave magnetron.

9. OPERATION OF THE TRAVELING-WAVE MAGNETRON

THE R-F PATTERNS OF THE MODES. In a magnetron of N segments there are N possible modes of oscillation. The different modes have periodicities of n, where n is any of the integers $0, 1, 2, \cdots N/2$ if N is even, and $0, 1, 2, \cdots, (N-1)/2$ if N is odd. A mode number, or designation, is numerically equal to n. Periodicity is here defined to mean the integral number of repeats of a field pattern in a single revolution around the anode at any given instant.

The tangential component of the r-f electric field is zero across the face of a segment. Therefore, the field distribution of a given mode will be the sum of an infinite series of harmonics. Each harmonic will have a periodicity of k. The only values of k possible are those for which

$$k = n - pN \tag{3}$$

where p is any integer—positive, negative, or zero. If the tube is of the rising-sun variety,

* The type III magnetron is also called the tangential resonance magnetron.

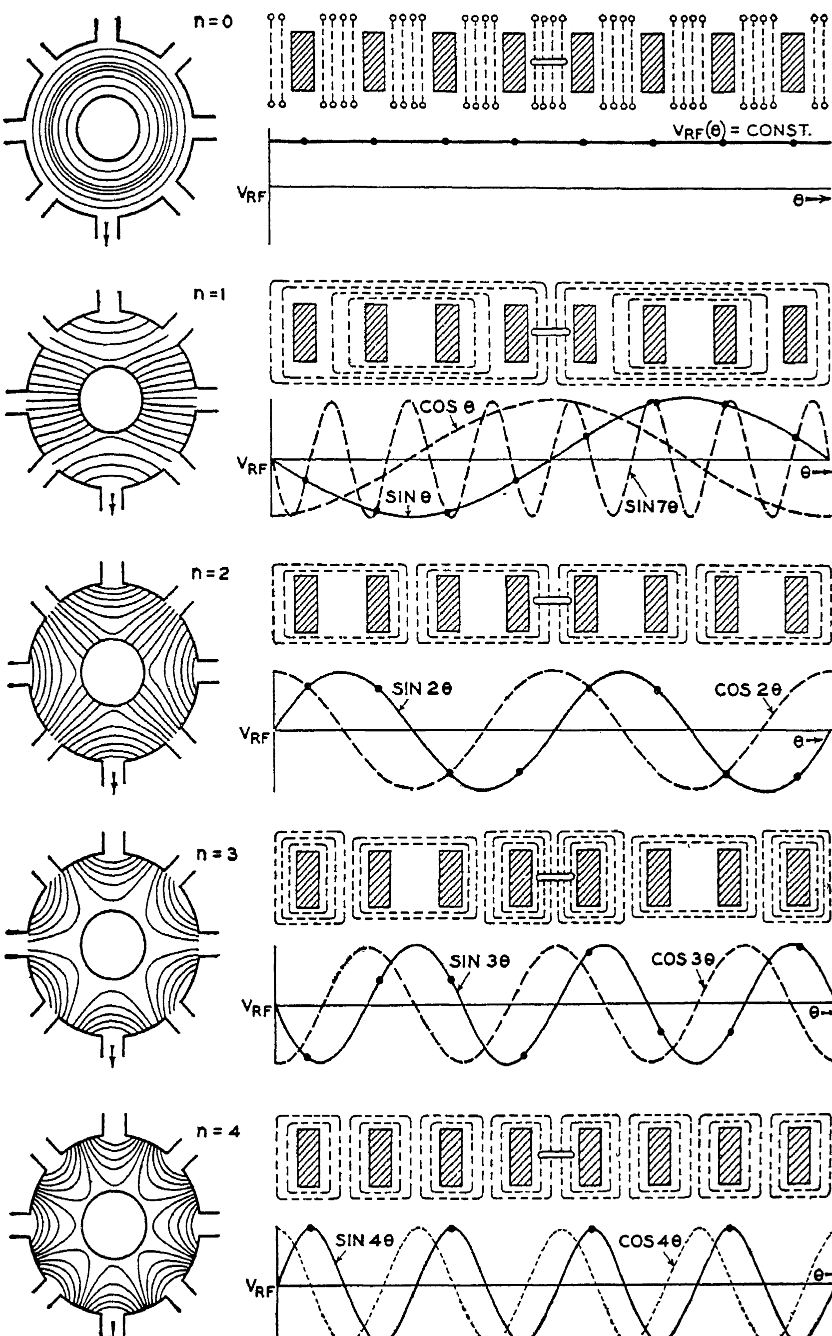


Fig. 3. Field and Potential Distributions of the Modes in an 8-segment Magnetron

where the resonators which connect the segments are alternately large and small, the allowed values of k will be given by

$$k = n - p\frac{N}{2} \tag{3a}$$

That is, there will be a set of harmonics associated with each of the two sets of resonators.

The mode for which $n = N/2$ is sometimes called the π mode for the reason that adjacent segments are 180°, or π radians, out of phase.

All the modes except the zero mode and the π mode occur in pairs, or doublets. The two members of a doublet have the same periodicity although they usually differ in frequency and the patterns are displaced with respect to one another in such a way that a current loop in the fundamental of one pattern occurs in the same position as a current node in the fundamental of the pattern of the other.

Some of these principles are illustrated in Fig. 3 for an eight-segment tube. For clarity only the electric flux lines of the fundamental component are shown in the interaction space. To illustrate the field configurations in the resonators, only the magnetic flux lines are shown. Below these are plotted the distributions in potential for the fundamental component. The reason for the existence of only one π-mode is illustrated in Fig. 3. The cos 4θ solution corresponds to zero potential on all the segments.

Although the frequencies of the several modes are usually different, all the harmonics of a given mode are at the same frequency. Thus, if $f_n(=\omega_n/2\pi)$ is the operating frequency of the nth mode, then the angular velocity of the pth harmonic of the rotating wave is ω_n/k, where k is given by eq. (3). This point is stressed for the reason that, in steady-state operation, the electronic stream interacts with only one harmonic of one mode at any instant.

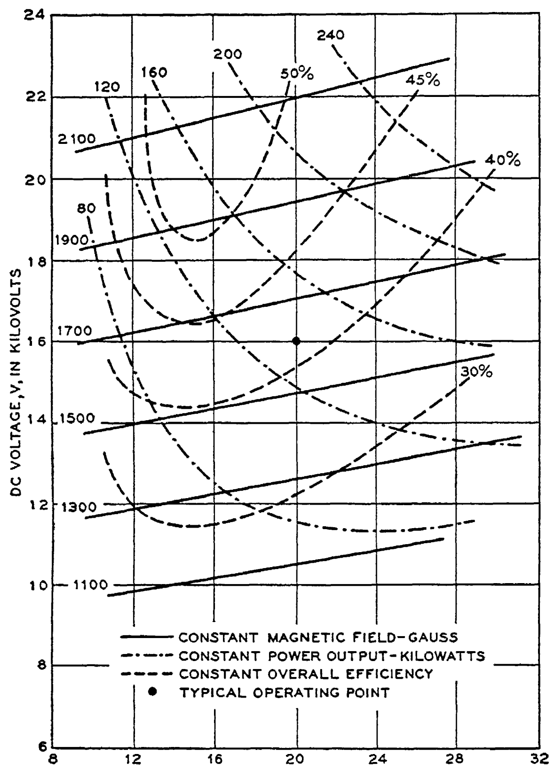


FIG. 4. Magnetron Input V-I Characteristic

INPUT CHARACTERISTICS. The d-c voltage for which an electron will just reach the anode for an infinitesimal r-f voltage on the anode of frequency f and periodicity k is given by

$$V_T = \frac{\pi f r_a^2 B}{k}\left[1 - \sigma^2 - \frac{2\pi f m}{keB}\right] \tag{4}$$

Equation (4) defines a threshold voltage. Empirically it is approximately equal to the operating voltage at low currents although changes in the r-f loading will cause changes in the input d-c voltage.

For any fixed loading the input voltage varies approximately linearly with input current for any fixed value of magnetic field, as is indicated in Fig. 4. Figure 4 shows voltage plotted as a function of current for various values of magnetic field. In addition to the constant magnetic field lines, contours of constant efficiency and constant power output are also shown. A method for obtaining the type of data required for such a V-I characteristic is described in Section 11, Microwave Measurements.

Although the performance chart in Fig. 4 is typical, it does not indicate the possible range of operating characteristics and parameters of magnetrons. Table 1 lists operating data for five magnetrons. The data for each tube are for some one point on its V-I characteristic. The most striking thing about the data is the large range of variation of the several parameters without any apparent correlation. It will be shown in the sequel, however, that the data do fit into a logical scheme.

Table 1

Example	Wavelength, centimeters	Voltage, volts	Current, amperes	Magnetic field, gausses	R-f Power Output, Watts
I	1	15,000	15.0	8,000	75,000
II	3	30,000	40.0	8,000	500,000
III	5	2,500	0.1	3,000	100
IV	10	50,000	200.1	2,500	4,000,000
V	50	2,000	1.0	800	1,000

SCALING. If, in the examples in Table 1, all the tubes were assumed to be working in an equivalent way (for example, all with the same electronic efficiency), then the data shown, together with the geometrical parameters, would correlate in accordance with the principles of scaling. The principles of scaling state that in the variation of the parameters, V, I, B, h, r_a, and λ, where h is the anode length, equivalent operation may be obtained by maintaining the following three parameters invariant:

$$V\left(\frac{\lambda}{r_a}\right)^2 \qquad I\left(\frac{\lambda}{r_a}\right)^2\left(\frac{\lambda}{h}\right) \qquad \lambda B \tag{5}$$

In addition, it is assumed that the ratio of cathode radius to anode radius and the number of segments also are maintained constant.

It would be instructive to consider an example of the type of scaling in which every linear dimension is scaled by the same factor, S. Let the unprimed quantities represent the tube from which it is desired to scale, and let the primed quantities represent the tube to be derived; then

$$\lambda' = S\lambda$$
$$r_a' = Sr_a$$
$$h' = Sh$$

From (5), since $(\lambda'/h') = (\lambda/h)$ and $(\lambda'/r_a') = (\lambda/r_a)$,

$$V' = V \qquad I' = I \qquad B' = \frac{B}{S}$$

Thus, the new tube will work at the same voltage and current but the new magnetic field will be $1/S$ times the original magnetic field. This type of scaling is most useful when the factor S does not differ greatly from unity. Ordinarily it cannot be applied if S is very large or very small. For example, suppose that $\lambda = 30$ cm and $\lambda' = 3$ cm. In this case, $S = 0.1$, and, since $I = I'$, the surface current density at the cathode will be up by a factor of 100 in the new tube. A current density of 10 amp per sq cm in the 30-cm tube is a quite moderate figure. A current density of 1000 amp per sq cm cannot be attained with present techniques. In order to scale over such wide ranges, it is usually necessary to extend the scaling laws to include N and σ. When these parameters are included, the quantities that must be kept invariant are

$$V\left(\frac{k\lambda}{r_a}\right)^2 \qquad I\left(\frac{k\lambda}{r_a}\right)^2\left(\frac{k\lambda}{h}\right)(1-\sigma^2)^2\left(\frac{1}{\sigma}+1\right) \qquad kB\lambda(1-\sigma^2) \tag{5a}$$

where k is used instead of N to emphasize the importance of the field periodicity. In π-mode operation k will be $N/2$.

Scaling in accordance with the invariants in (5) and at the same time maintaining the same N and σ gives fairly accurate results. Scaling in accordance with the invariants in (5a) will give good results in predicting currents, voltages, and magnetic fields. However, it is usually found that equivalence does not hold for large changes in N or in σ. In particular it is usually found that electronic efficiency decreases with increasing N and increasing σ.

MODE SEPARATION AND MODING. The frequencies of the several modes are separated by an amount which depends upon the type and degree of coupling among the several resonators. Such mode separation is advantageous for the reason that it allows

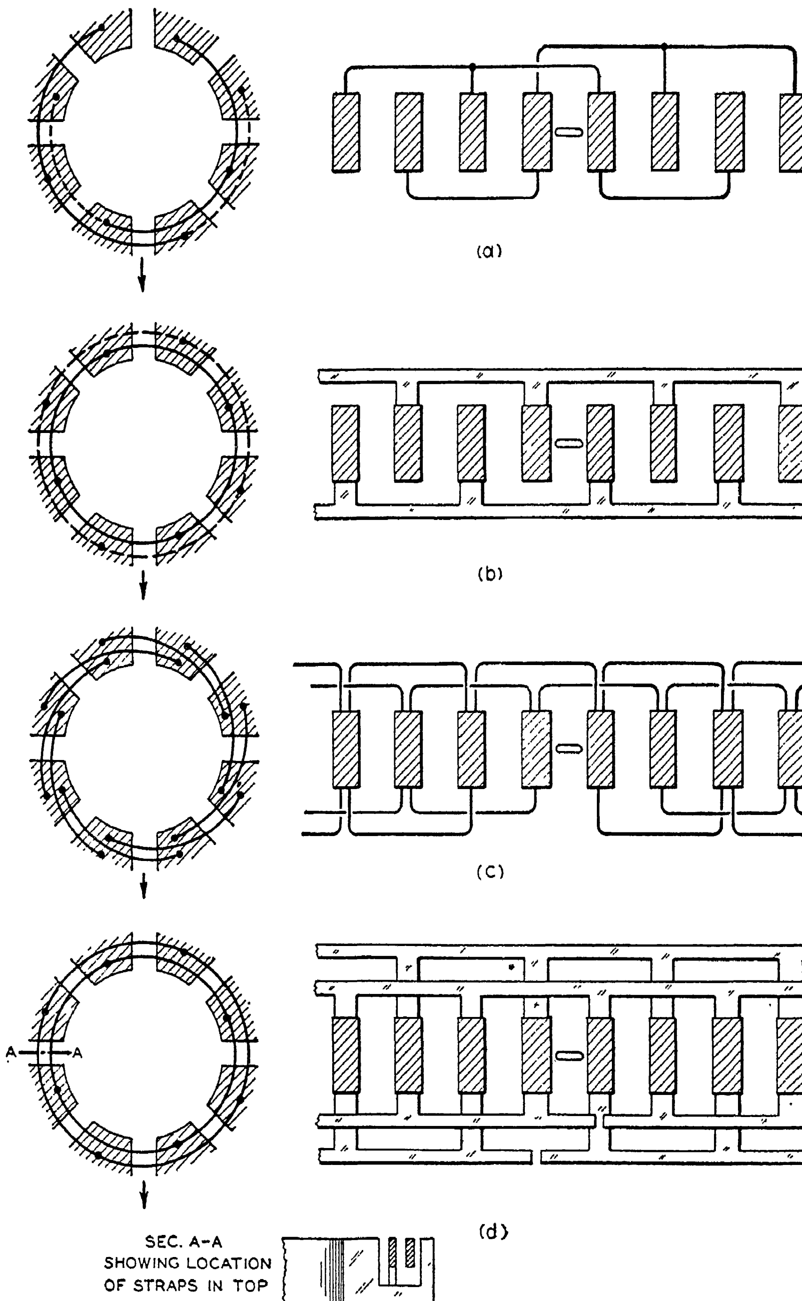


FIG. 5. Magnetron Anode Strapping Methods. (*a*) Early British, (*b*) single ring, (*c*) echelon, (*d*) double ring.

the desired mode of operation to be excited independently of the other modes. This appears to be one of the conditions necessary for high electronic efficiency.

From eq. (3) it might also seem desirable to separate the values of f/k in order that no two modes have the same threshold voltages. This has appeared to be true in some cases, but, in general, when taken alone, this condition is an unreliable index of the possibility of moding. The problem is complicated by several factors including, principally, such things as relative r-f loading of the modes, the noise levels at the start of oscillations, the transient behavior of the modulator and power supply, and the instability of the space charge of the magnetron at high current levels. A complete analysis of moding is beyond the scope of this article. However, a few general remarks can be made. In these remarks, it will be assumed that N is even and the π-mode operation is desired.

Two distinct types of moding have been observed. One, called the mode skip, occurs principally in the high-voltage pulse magnetrons. The magnetron fires in both the π mode and an unwanted mode. ("Unwanted mode" is to be considered as being defined here so as to include a non-oscillating state which sometimes occurs.) However, it fires in only one mode during any one pulse. It alternates between the two modes in more or less random fashion. This type of moding can usually be cured either by reducing the r-f loading on the magnetron, by reducing the rate of rise of the applied voltage pulse or by reducing the applied voltage, or by a combination of the three.

A mode shift, the other type of moding, is encountered chiefly in low-voltage c-w magnetrons. It consists of a shift from one mode to another. One of its principal causes is an inherent instability in the space charge at high current. In general, it can be cured by reducing the r-f loading or by reducing the operating current. In the pulsed case, a mode shift is usually observed to occur during a single pulse and is relatively unaffected by changes in the rate of rise of the applied pulse. If the cathode is well designed, the high current instability will occur at currents which are lower than those necessary for temperature limitation. Occasionally, however, mode shifts have been observed which involve instability due to temperature-limited operation.

METHODS OF MODE SEPARATION. The two most important devices for achieving frequency separation of the modes involve the use of straps and of the rising-sun structure.

Figure 5 shows several strapping methods schematically. One of the most widely used is the double ring strapping of Fig. 5*a*, in which there is a pair of concentric rings at each end of the anode. One ring of a pair is connected to one set of alternate segments. The other ring is connected to the other alternate set of segments. In π-mode operation, one strap or ring is everywhere at the same potential and the two rings at one end are 180° out of phase. Hence, the straps constitute mainly a capacitance loading of the resonators with a resultant increase in wavelength in the π mode over the unstrapped case. In the $n = 1$ mode, the potential distribution on a strap is almost sinusoidal and periodic in only one revolution around the anode. Moreover, the two rings in a set are at nearly the same potential. Thus, the principal effect of straps in the $n = 1$ mode is a shunt inductance so that the $n = 1$ mode wavelength will be less in the strapped case than it is in the unstrapped case. Modes in between the $n = 1$ and the π mode will be affected in a way intermediate between the two extremes first mentioned. That is, as n increases from 1 to $N/2$, the inductance effect of the straps decrease while the capacitive effect increases.

Figure 6 shows the mode spectra of three different types of 18 segment magnetrons. The spectrum in Fig. 6*a* is for an unstrapped symmetrical tube, that in Fig. 6*b* for a strapped symmetrical tube, while Fig. 6*c* shows the spectrum for a rising-sun tube.

At wavelengths shorter than 3 cm, the mechanical problem of making strapped magnetrons becomes very difficult. The rising-sun structure avoids this difficulty by providing mode separation without straps. It works on the principle that the normal mode frequencies of a system of resonators become separated by the introduction of asymmetries. The mode spectrum of the rising sun is composed of two branches. Each branch is associated with one of the two sets of resonators.

OUTPUT COUPLING. The r-f energy can be coupled to the load in several ways. Of these, two are of major interest: loop coupling and wave-guide coupling. The first type is indicated schematically in Fig. 2*a*; the second is shown schematically in Fig. 7. In Fig. 7, coupling is obtained through a slot in the back of one of the resonators. The quarter-wave low-impedance transformer section steps down the impedance of the wave guide, which is of the order of a few hundred ohms, to the low impedance required at the slot opening, which is, in some cases, as low as 1 or 2 ohms.

The choice between loop output or wave-guide output is usually made on the grounds of mechanical feasibility and convenience. At wavelengths greater than 10 cm, r-f transmission is ordinarily done in coaxial lines. The size of choke assemblies and window structures in wave guides makes them cumbersome and difficult to fabricate. For these rea-

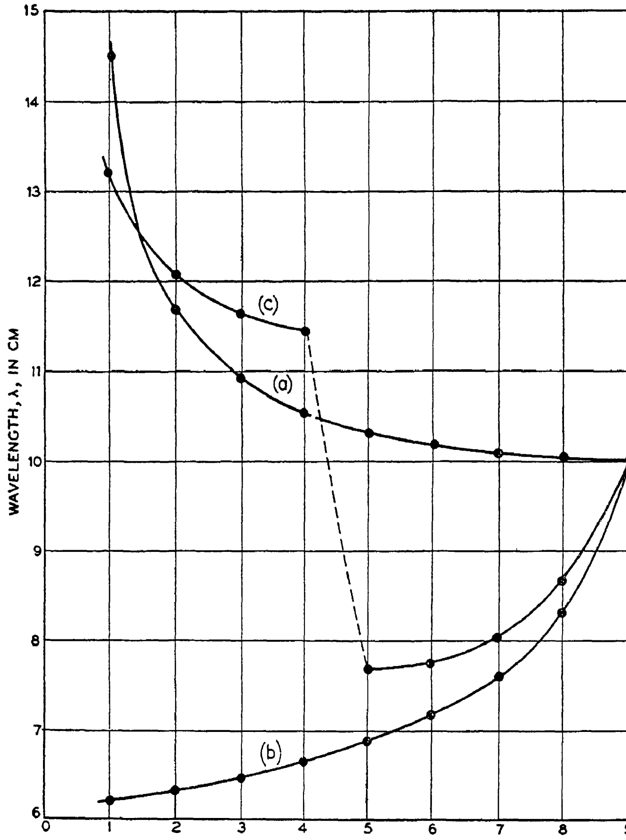


FIG. 6. Mode Spectra for Three Different Types of 18-segment Magnetrons Having the Same π-mode Wavelength. (*a*) Unstrapped symmetrical magnetron, (*b*) strapped symmetrical magnetron, (*c*) rising sun magnetron.

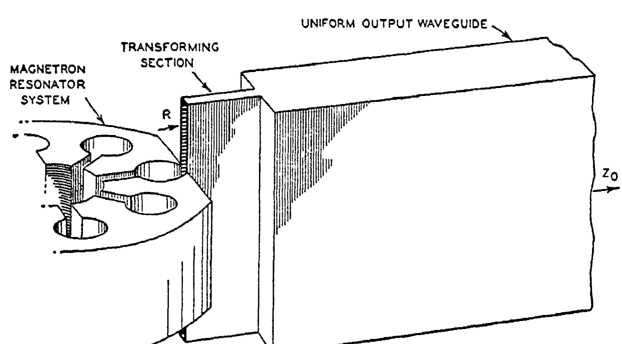


FIG. 7. Illustration of a Method of Wave Guide Output Coupling

sons, loop coupling outputs are usually found at the longer wavelengths. Below 10 cm, wave guides are usually used for transmission, and the physical size of wave-guide outputs is reduced to manageable proportions.

In common with any self-excited oscillator, the efficiency and frequency stability of a magnetron depends upon the amount of coupling between the magnetron and the load. In addition, for any given coupling, the efficiency and frequency stability will change as the amount of loading is changed. A quantitative measure of the variation of efficiency and frequency with loading is obtained from a Rieke diagram, an example of which is shown in Fig. 8. Contours of constant power output and contours of constant frequency are

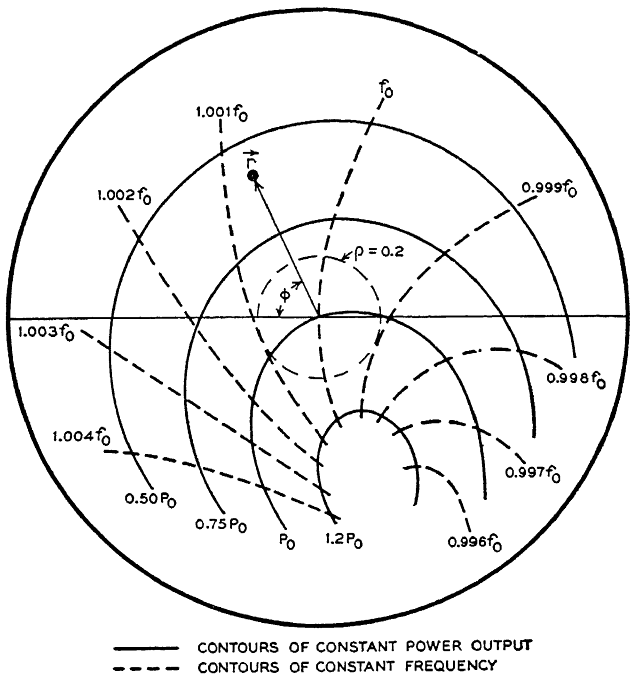


FIG. 8. Rieke Diagram

plotted on a polar diagram. Load impedances on a polar diagram may be specified by a characteristic impedance of the transmission line into which the magnetron is coupled, together with a reflection coefficient. The modulus of the reflection coefficient is proportional to the distance from the pole of the diagram while its argument is proportional to the angular displacement around the diagram. The load line characteristic impedance Z_0, the load impedance Z, and the voltage reflection coefficient r are related by

$$r = \frac{Z - Z_0}{Z + Z_0} = \rho e^{j\phi} \tag{6}$$

where ρ is the modulus and ϕ is the argument of r. The region on the Rieke diagram of high power and high frequency density is the region of heavy loading. The center of the diagram corresponds to a matched load; that is, the load impedance is equal to the transmission-line characteristic impedance. Ordinarily, the loading is maintained constant, at the match point, over the V-I characteristic, as in Fig. 4. On the other hand, the magnetic field and either the current or voltage is held constant over a Rieke diagram. A description of some of the techniques for obtaining the data for a Rieke diagram and a discussion of polar diagrams will be found in Section 11, Microwave Measurements.

A quantitative measure of the frequency stability with respect to perturbations in loading at the match point is the pulling figure. Pulling figure is defined as the greatest excursion of frequency observed as the modulus of the reflection coefficient is maintained constant at a value of 0.2 and its phase angle is varied through 360°. For the example shown, the pulling figure is about 0.2 per cent of the center frequency. For a magnetron operating at a wavelength of 3 cm this would represent a pulling figure of 20 Mc.

If the coupling from the magnetron to the transmission line is made tighter, the power output and the pulling figure at the match point will both increase. Nearly all magnetron output circuits are designed to provide some predetermined compromise of efficiency and pulling figure when the tube is operated into a matched line.

TUNING. The frequency of oscillation is determined almost entirely by the geometry of the cavities, that is, by the equivalent inductance and capacitance of the oscillator tank circuit. In order to tune the magnetron, it is necessary to change either the inductance or capacitance or both.

Figure 9 shows three tuning methods schematically. The inductive pin tuning method is depicted in (*a*). An array of copper pins is disposed so that they may be inserted and

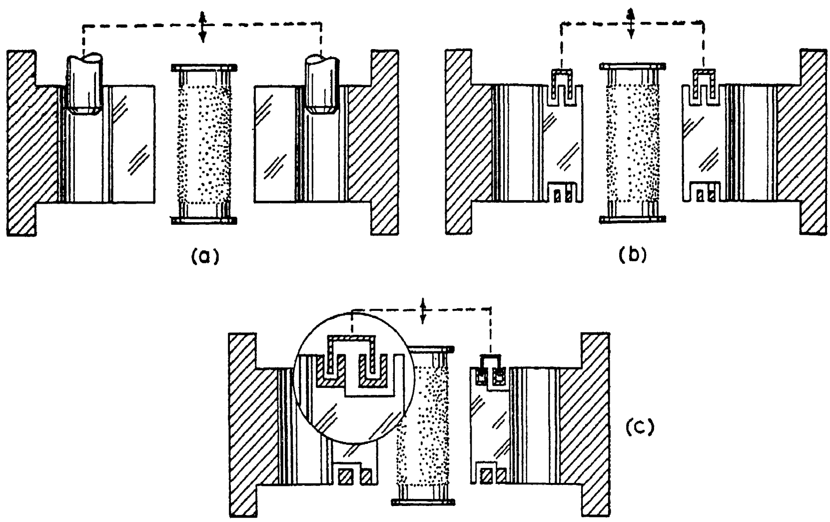


FIG. 9. Methods of Magnetron Tuning. (*a*) Inductive pin, (*b*) segment-to-segment capacitance, (*c*) strap-to-strap capacitance.

withdrawn from the inductive portion of the resonators. As the pins are inserted, the effective inductance is decreased and the frequency goes up. Capacitance variation schemes are shown in Figs. 9*b* and 9*c*. In (*b*), a movable conducting ring changes capacitance between the segments. In (*c*), the ring changes the capacitance of the straps. The schemes shown in (*a*) and (*c*) are capable of giving tuning ranges of about 20 per cent. The scheme shown in (*b*) is usually limited to a somewhat shorter tuning range for the reason that different modes tune at such widely varying rates with respect to tuner displacement that the π mode tunes but a relatively short distance before encountering interference from other modes.

BIBLIOGRAPHY

Collins, G. B. (ed.), *Microwave Magnetrons*, M.I.T. Radiation Laboratory (sponsor). McGraw-Hill (1947).
Brainerd, J. G., *et al.*, *Ultra-High Frequency Techniques*. Van Nostrand (1942).
Fisk, Hagstrum, and Hartman, *Bell Sys. Tech. J.*, Vol. 25, 167 (1946).
Herriger and Hülster, *Zeit. f. Hochfrequenz.*, Vol. 49, 123 (1937).
Kilgore, G. R., *Proc. I.R.E.*, Vol. 24, 1140 (1936).
Posthumus, K., *Wireless Engineer*, Vol. 12, 126 (1935).
Megaw, E. C. S., *J. I.E.E. (London)*, Vol. 72,326 (1933).
Okabe, K., *Proc. I.R.E.*, Vol. 17, 652 (1929).
Hull, A. W., *Phys. Rev.*, Vol. 18, 31 (1921).
Habann, *Zeit. f. Hochfrequenz.*, Vol. 24, 115 and 135 (1924).
Blewett and Ramo, *Phys. Rev.*, Vol. 57, 635 (1940).

KLYSTRONS

By A. L. Samuel

Electron tubes which make use of the principle of velocity modulation are now known as klystrons. The name klystron was originally a registered trademark. It is now applied to all tubes of the same general type without regard to the manufacturer. A klystron has been defined by the IRE as an electron tube in which the distinguishing features are the modulation or periodic variation of the longitudinal velocity of an electron stream without appreciable variation of its convection current and the subsequent conversion of this velocity modulation into convection-current modulation by the process of punching. All commercial tubes of the klystron type make use of cavity resonators, although such resonators are not, in principle, essential to their operation. Two types of klystrons are in general use: tubes of the first type bear no further designation: tubes of the second type are referred to as reflex klystrons or simply reflex tubes.

10. KLYSTRONS (EMPLOYING TRANSIT TIME BUNCHING)

A few typical klystrons are shown in Figs. 1 and 2. As already stated, it is customary, although not essential, to employ resonant cavities as the tuned circuits associated with the input and output portions of these amplifiers. These cavities take the place of conventional circuits and must be tuned to the operating frequency. They may be partly external to the tube proper, or they may form an integral part of the tube as supplied to the user. Cavities are used because they produce larger effective fields in the interaction gap regions than could be obtained by any other means.

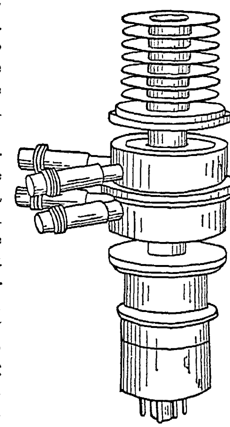


FIG. 1a. A Typical Klystron of the Integral Cavity Type Shown without Its Tuner. (The 3K30 oscillator amplifier.)

The basic principles of the klystron amplifier may be explained by referring to Fig. 3. This figure illustrates a tube which consists of (1) an electron gun, composed of a heater, a cathode, and auxiliary focusing electrodes; (2) an input region called the input gap, defined by two grids which in this case form a part of a cavity resonator; (3) a conversion region called the drift space which is relatively free of electric or electromagnetic fields; (4) an output region, called the output gap, again defined by two grids which form a part of a second cavity resonator; and (5) a collector electrode whose sole function is to collect the electron stream after it has traversed the working region of the tube. These five portions of the tube correspond directly to the five essential operations that must be performed in any electron tube. The separation of these operations makes it possible to consider them separately and to explain the operation of the device in very simple terms. These operations are, obviously, (1) the production of an electron stream, (2) the modulation or variation of some property of this stream in accordance with an input signal, (3) the conversion of the original modulation into a form in which it can be utilized, (4) the utilization of the stream to produce an output signal, and (5) the collection of the electron stream.

Referring to Fig. 3, the field in the input gap region of the tube varies the velocity of the electron stream in a cyclic manner, the variation in velocity being assumed to be small compared to the average velocity imparted to the stream by the d-c fields. Those electrons which arrive when the field is in an aiding direction are speeded up; those arriving a half cycle later are slowed down. The contributions in energy made by the field to some of the electrons of the stream is nearly balanced by the energy taken from those electrons which are slowed down. The modulating process, therefore, requires substantially no power, most of the input power being consumed by ohmic losses in the walls of the input cavity. All the electrons, except those intercepted by the grids, proceed through the next region of the tube, the so-called drift space, where the electron stream becomes bunched through the simple process of the faster electrons, that is, those that are speeded up by the field in the input gap, overtaking the group of slower electrons that precede them. This bunching process converts the original velocity modulation into a variation in the rate at which electrons pass any given point. The stream as it crosses the output gap appears superficially like the stream of electrons in the screen-grid plate region of the conventional space-

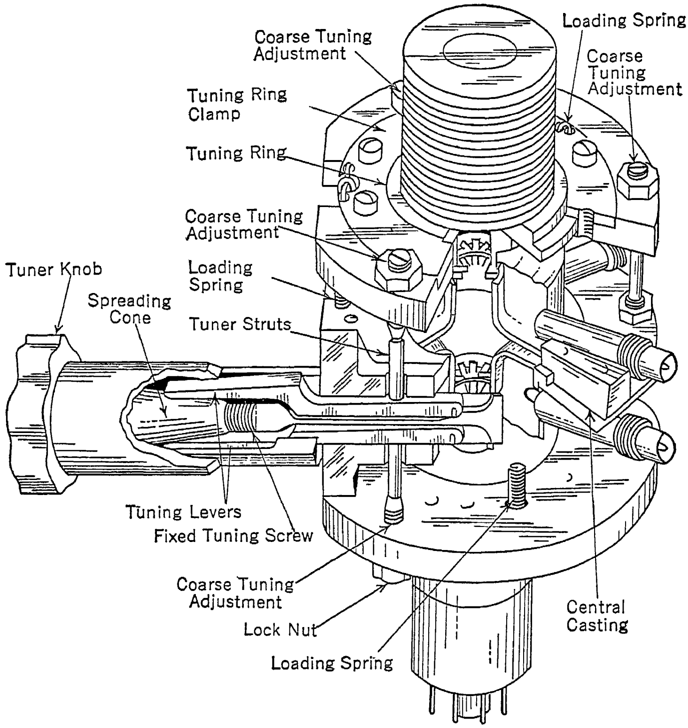


FIG. 1b. The 3K30 Shown in Section with an 11-C Tuner Attached

charge control tetrode. It, therefore, induces currents in the output cavity and delivers power to the output in just the same way that an electron stream delivers energy to the field between the screen and the plate of the conventional tube. The tuning of the output cavity must be adjusted to cause the maximum value of the field to occur in a retarding direction at the time that the effective center of an electron bunch crosses the output gap. Finally, the spent electron stream is collected by a final electrode where the energy remaining in the stream is dissipated as heat. This is to be contrasted with the action in the conventional tube where the plate performs the dual function of providing the output circuit retarding field and of dissipating unused energy as heat.

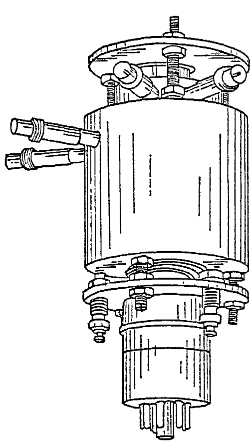


FIG. 2. The Type 2K47 Klystron Frequency Multiplier

An interesting possibility exists in the klystron amplifier of utilizing the electron stream in cascade to provide either a multistage amplifier or a combination function device such, for example, as an oscillator buffer amplifier. When the electron stream traverses the output gap in the simple amplifier, it obtains an augmented velocity modulation as a result of the higher field intensity existing in this cavity—this at the same time that the stream delivers energy to the cavity because of its bunched condition. This augmented modulation is in quadrature with the original

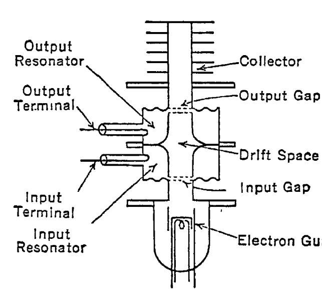


FIG. 3. A Sectional View of a Klystron Amplifier

modulation. At low modulation levels, such as those obtained in an amplifier, it can be thought of as existing quite independently of the original modulation. In the ordinary two-cavity single-stage amplifier, no use is made of this added modulation. However, by providing a second drift space and a third cavity, an additional stage of amplification can be obtained. The middle cavity or cascade cavity need have no external connection, although, if broad band amplification is desired, it is necessary to load this cavity in some fashion to reduce its effective Q to a value comparable to that of the input and output cavities which are loaded by their external circuit connections.

The gain of a klystron amplifier varies with the input signal level in quite a different way from the behavior of other types of amplifiers. At low levels, the gain depends only on the beam current, the beam voltage, and the physical dimensions of the tube. However, as the drive is increased beyond a certain point, the phenomenon of overbunching sets in and the gain begins to decrease. Eventually a maximum output is reached; with a further increase in the input, the output actually decreases. This is illustrated in Fig. 4, where the output as a function of the input is plotted for a typical klystron amplifier.

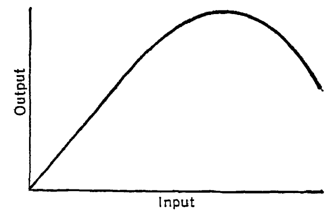


FIG. 4. The Variation in Output Power with the Input Showing the Effect of Overbunching

At low levels, the gain of a klystron amplifier is given by

$$\text{Gain} = Z_1 Z_2 M_1^2 M_2^2 S^2 \tag{1}$$

where Z_1 and Z_2 are the impedances of the input and output cavities respectively as measured across the interaction gaps. The parameters M_1 and M_2 are the absolute values of the input and output beam coupling coefficients or so-called modulation coefficients and express the effectiveness with which the fields in the cavities interact with the electron stream. The parameter S is the absolute value of the beam transadmittance and is a measure of the ratio of current variation produced by the bunching process to the equivalent voltage variation impressed on the beam.

The cavity impedances can be measured directly if desired, or they may be computed from measured values of the characteristic impedance and Q of the cavity.

The absolute value of the beam coupling coefficient for the gap between two axial cylinders without grids varies for different electrons depending upon their distance from the axis of the beam. The value for that portion of the beam lying at a radial distance from the axis of $\theta\rho$ where $\theta\rho$ is expressed in radians at the operating frequency and with tubes having a radius of θ_r (also in radians) is given by

$$M = \frac{\sin(\theta_g/2)}{\theta_g/2} \cdot \frac{I_0(\theta\rho)}{I_0(\theta_r)} \tag{2}$$

where the I_0's are modified Bessel functions of the first kind.

The absolute value of the beam coupling coefficient for an interaction gap between ideal grids is given by

$$M = \frac{\sin(\theta_g/2)}{\theta_g/2} \tag{3}$$

where θ_g is the electron transit angle of the gap defined as the time required for the unmodulated electron beam to cross the gap measured in radians at the operating frequency. The value of θ may be compiled from

$$\theta_g = \frac{3180X}{\lambda\sqrt{V}} \tag{4}$$

where X is the gap spacing, λ is the free space wavelength corresponding to the operating frequency, and V is the voltage corresponding to the velocity of the electron beam. The parameters X and λ are measured in the same units (usually centimeters), and V is in volts.

The absolute value of the beam transadmittance for low signal levels is given by

$$S = \frac{\theta I}{2V_0}\sigma \tag{5}$$

where θ is the electron transit angle in the drift space, I is the beam current in amperes, V_0 is the beam voltage, and σ is a reduction factor to account for certain space-charge effects that will not be discussed. In well-designed amplifiers, the parameter σ is usually of the order of 0.5 at the recommended operating conditions and increases to 1.0 as the beam current (at a given voltage) is decreased to a low value.

At high levels the beam transconductance is sometimes expressed as

$$S = 2IJ_1 \cdot \frac{(V\theta)}{(2V_0)} \tag{6}$$

where V is the value in volts of the velocity modulation impressed on the beam, and J_1 is a Bessel function of the first kind.

This expression is based on kinematic considerations only and neglects space charge and other sources of non-linearity. It may be used as a rough basis for predicting the general behavior of a klystron amplifier for large signals, but it does not agree quantitatively with experimental results.

11. REFLEX KLYSTRONS

A reflex tube or reflex klystron is a special form of the klystron oscillator employing a single cavity with a single interaction gap to perform the functions of both the input and output circuits (see Figs. 5 and 6). The electron stream is velocity modulated on a first transit of this gap and is forced to cross the gap a second time by means of a repelling or reflecting field. The electrons become bunched in the process of reflection, the speeded-up electrons penetrating the field to a great distance and therefore taking longer to return than the slowed-down electrons in just the same way that a ball thrown upward in the earth's gravitational field takes longer to return if thrown with high velocity than if thrown with low velocity. In order for the oscillations to be self-sustained, the returning bunches of electrons must arrive at the interaction gap at the correct phase of the alternating field, that is, when the field has its maximum value in the retarding direction. This requires that the electrons remain in the reflecting field region for a critically valued length of time, a time that is approximately $n + 3/4$ cycles at the operating frequency, where n is any integer equal to or greater than zero. This time can be adjusted by varying the velocity of the beam as it crosses the gap on the first transit or, more usually, by varying the voltage of the repeller. As the voltage of the repeller is varied, a series of operating regions

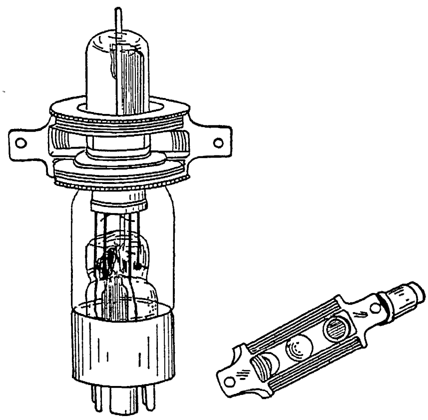


FIG. 5. The 707A Reflex Klystron Employing an External Cavity, Shown with the Cavity Partly Disassembled

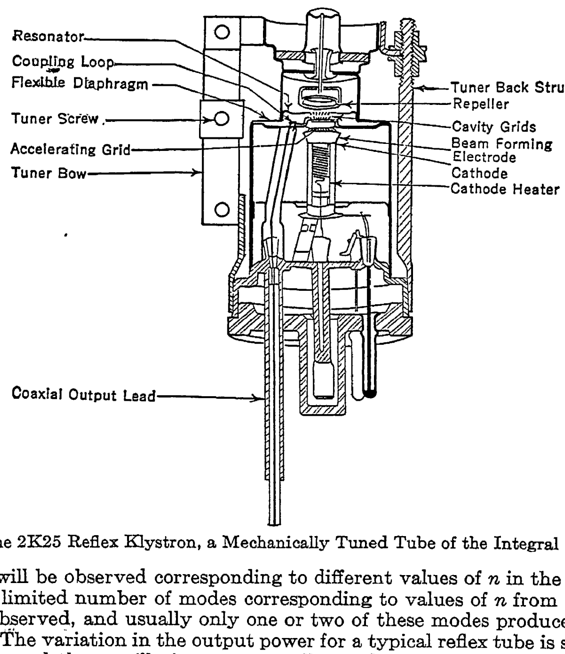


FIG. 6. The 2K25 Reflex Klystron, a Mechanically Tuned Tube of the Integral Cavity Type

called modes will be observed corresponding to different values of n in the above relationship. Only a limited number of modes corresponding to values of n from 1 or 2 to 4 or 5 are actually observed, and usually only one or two of these modes produce enough power to be useful. The variation in the output power for a typical reflex tube is shown in Fig. 7. It will be observed that oscillations are actually produced for a limited range in voltage in the vicinity of the optimum values, and that a variation of frequency occurs as shown by the top curves. This variation in frequency with voltage is called electronic tuning.

Electronic tuning is often employed as a means for critically adjusting the operating frequency in applications where the accompanying variations in the output power can be tolerated, such, for example, as a local oscillator in a superheterodyne receiver. Electronic tuning finds its greatest usefulness in connection with automatic frequency control circuits. Since the mode with the highest output has the smallest electronic tuning band width, a compromise must often be made between output and tuning range. The electronic tuning range is generally small as compared to the usual mechanical tuning range. When electronic tuning is employed, it is essential that the mechanical tuning be adjusted so that the operating point falls somewhere near the middle of the electronic tuning range where the frequency can be shifted a reasonable amount in either direction with the electronic control without too much change in output power. Electronic tuning and other associated phenomena can be explained by considering the way in which the impedance of the electron stream as seen by the cavity varies with the electron transit time in the repeller region.

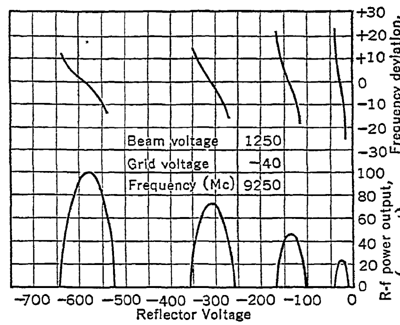


Fig. 7. A Typical Mode Curve for a Reflex Klystron Showing the Variation in Output Power and Frequency with the Reflector Voltage (Data for the 2K41 Tube)

Figure 8 is a plot of the small signal beam conductance showing the relationship

$$Y_g = \frac{M^2\theta I}{2V_0}\frac{J_1(VM\theta/2V_0)}{VM\theta/V_0}e^{j(\pi/2-\theta)} \tag{7}$$

where the first term represents the magnitude of small signal admittance, and the last term is the phase angle. Here θ is the transit angle in the repeller region, and the rest of the symbols have the same significance as in the article on klystrons employing transit time bunching. The second term accounts for the decrease in magnitude of the conductance which occurs for large signal levels and is similar to the compression term appearing in the expression for the small signal transadmittance of the klystron amplifier. The final term gives the phase angle of the conductance.

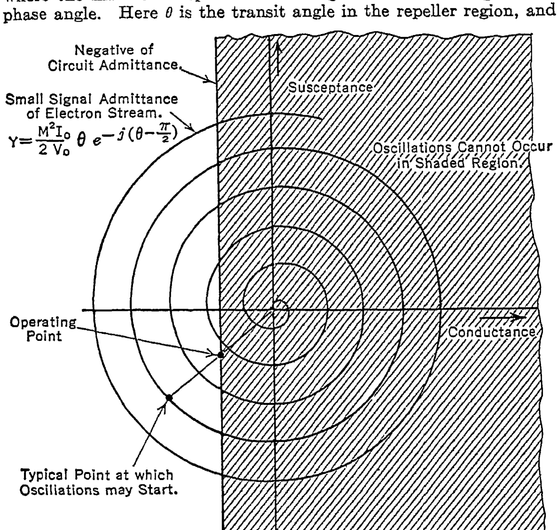


Fig. 8. The Admittance of the Electron Stream as Viewed from the Cavity for a Reflex Klystron. Oscillations can occur only in the region where the negative conductance of the beam exceeds the positive conductance of the circuit.

If the real part of this admittance is negative and larger in magnitude than the positive real component of the cavity admittance, oscillations will build up until limited by the non-linearity given by the second term. Under stable operating conditions, the relationship

$$Y_g + Y_c = 0 \tag{8}$$

must be satisfied, where Y_c is the admittance of the cavity. The cavity may be assumed to behave as a simple shunt tuned circuit in the vicinity of the resonant frequency so that

$$Y_c = G + \frac{j}{Q_0}\left(\frac{\omega}{\omega_0} + \frac{\omega_0}{\omega}\right) \tag{9}$$

where G is the cavity conductance, Q_0 is the cavity Q, ω_0 is the angular frequency at resonance, and ω is the angular frequency corresponding to the particular value of Y_c. The negative of this value is plotted in Fig. 8 and appears as the straight line to the left of the imaginary axis.

Oscillations will not be sustained for all values of Y_g lying in the shaded area on the figure. For values of Y_g lying to the left of the Y_c line, oscillations will build up until Y_g sinks along a radial line arriving at a stable operating point on this line. If the operating point lies on the real axis, the oscillations will occur at the resonant frequency of the cavity. Operating points off the axis correspond to oscillations at a frequency which differs from the resonant frequency by a sufficient amount to provide the necessary reaction component of admittance specified by the operating point. This effect is called electronic tuning.

The output impedance characteristic of a reflex tube can best be illustrated by plotting this characteristic on the reflection coefficient plane (sometimes called a Rieke diagram on a Smith chart). A typical plot is shown in Fig. 9, where lines for constant power are shown solid and lines for constant frequency are plotted on a background of orthogonal circles representing fixed values of the resistive and reactive components of the load impedance. The region on the plot where the constant-frequency lines tend to converge is called the

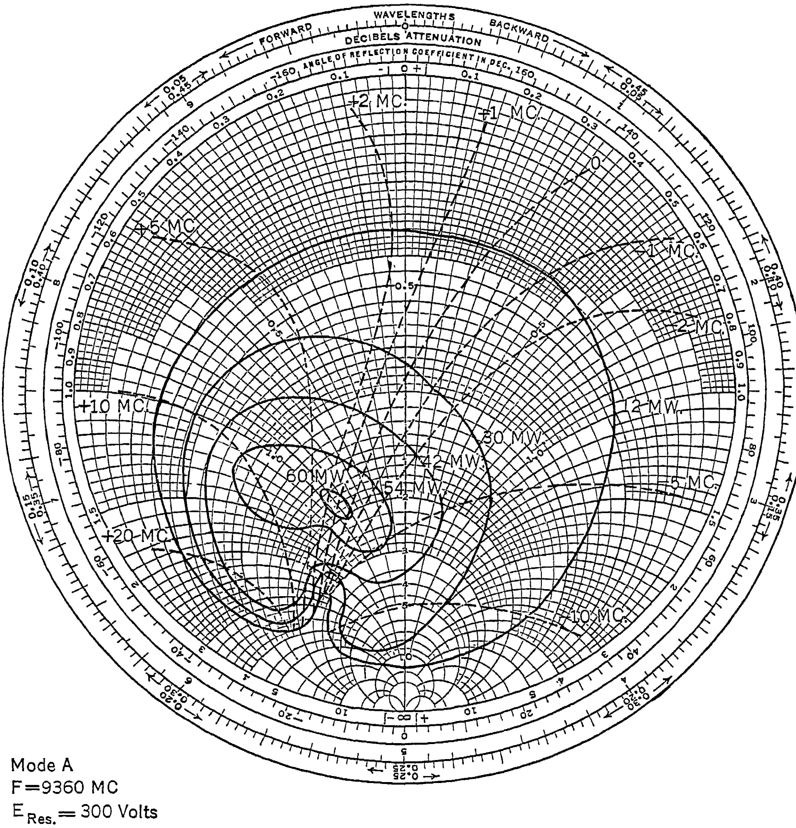


FIG. 9. The Variation in Output Power and Frequency with the Load Impedance for a Typical Reflex Klystron (the 2K25) Shown on the Reflection Coefficient Plane

frequency "sink," and the minimum amplitude of standing wave ratio that will cause the tube to operate in this region of discontinuity is called the "sink margin." It is customary to require a sink margin of 8 db. A second important characteristic is the so-called pulling figure which is defined as the maximum difference in frequency produced when a mismatch, having a reflection coefficient of 0.2 as measured at the prescribed output coupler, is varied through 360°. The pulling figure for the tube shown in Fig. 9 is approximately 4 Mc.

The power output, electronic tuning sink margin, and pulling figure for the typical reflex tube will be found to vary somewhat over the mechanical or thermal tuning range of the tube. Curves showing these variations for any particular tube are customarily supplied in the technical information sheets published by the manufacturers.

The coarse adjustment of frequency is usually made by mechanical means, either by varying the effective size of the external portion of the cavity or by internal changes, usually of the length of the interaction gap, and hence of the effective capacitance loading of the cavity. In most tubes where capacitance tuning is employed, the necessary motion is transmitted through the vacuum envelope by means of a flexible diaphragm. Recently, a number of tubes have been introduced in which mechanical cavity tuning is produced internally by thermal expansion means. This method, called thermal tuning, makes it possible to adjust the frequency of the tube over its entire mechanical tuning range by electrical means, without the large variations in output power that are encountered with electronic tuning. The thermal tuning speed is limited by the thermal capacity of the tuning mechanism, but it is sufficiently fast for many automatic frequency control applications.

12. TUBE TYPES

Many of the tubes listed in Table 1 were made for the armed services during World War II, and some may not be commercially available. The prospective user should consult the manufacturer in regard to their availability and should follow his recommendations regarding operating conditions and ratings. The data of Table 1 are indicative of typical operating conditions and are supplied for general reference purposes only.

Table 1. Western Electric Reflex Tubes—External Cavity Type

Number	Frequency, megacycles	Resonator Voltage	Reflector Voltage (Negative)	Heater Voltage	Output, milliwatts	Tuning Range, megacycles			Remarks
						Mechanical	Electronic	Thermal	
707A/B	2,500– 3,750	300	0– 275	6.3	70	*	35	....	
2K48	3,000–10,000	1,000	0– 500	6.3	24	*	10	......	

Table 1—*Continued.* Western Electric Reflex Tubes—Internal Cavity Type

Number	Frequency, megacycles	Resonator Voltage	Reflector Voltage (Negative)	Heater Voltage	Output, milliwatts	Mechanical	Electronic	Thermal	Remarks
726C	2,700– 2,930	300	50– 210	6.3	120	230	25	....	
726B	2,880– 3,170	300	50– 210	6.3	120	290	25	....	
726A	3,170– 3,410	300	50– 210	6.3	120	240	25	....	
2K29	3,400– 3,960	300	50– 210	6.3	95	560	32	....	
2K56	3,840– 4,460	300	125– 175	6.3	65	620	30	....	
2K54	4,290– 4,560	1,130	800–1,100	6.3	700 †	270	......	....	Pulsed
2K23	4,275– 4,875	1,130	600– 900	6.3	250 †	600	......	....	Pulsed
2K55	4,590– 4,860	1,130	625– 850	6.3	700 †	270	.....	....	*Pulsed*
2K22	4,240– 4,910	300	75– 235	6.3	85	670	35	....	
2K26	6,250– 7,060	300	70– 150	6.3	50	810	35		
2K25	8,500– 9,660	300	75– 200	6.3	30	1,160	32		
2K45	8,500– 9,660	300	95– 145	6.3	30	....	50	1,160	8 sec tuning time
723A/B	8,700– 9,550	300	90– 200	6.3	25	850	32		
2K50	23,215–24,750	300	20– 130	6.3	10	.. ..	65	935	2 sec tuning time

The data above were obtained on tubes manufactured for the Army and Navy.

Table 1—*Continued.* Raytheon Manufacturing Company

Type	Frequency, megacycles	Resonator Voltage	Reflector Voltage (Negative)	Heater Voltage	Output Power, milliwatts	Tuning Range, megacycles			Remarks
						Mechanical	Electronic	Thermal	
QK269	1,200– 1,500	300	100–220	6.3	150	300	12		
707B	3,400– 3,600 *	300	155–290	6.3	150		20		Requires external cavity
2K28	3,400– 3,600 *	300	155–290	6.3	150		20		Requires external cavity
QK159	2,950– 3,250	300	112–250	6.3	150		20		
5721	4,290– 8,340 *	1000	60–600	6.3	160		12 (min.)		Requires external cavity
2K25/723A-B	8,500– 9,660	300	85–200	6.3	33	1160	40		
2K33	23,710–24,290	1800	80–220	6.3	40	580	40		

Table 1—*Continued.* Sylvania Electric Products, Inc.

Type	Frequency, megacycles	Resonator Voltage	Reflector Voltage (Negative)	Heater Voltage	Output Power, milliwatts	Mechanical	Electronic	Thermal	Remarks
6BL6	1,600– 5,500 *	350	15–700	6.3	125		10		Requires external cavity
6BM6	500– 3,000 *	350	15–700	6.3	60		13.4		Requires external cavity

* Tuning range of external cavity tubes depends upon cavity design and may be anything up to the total range of the tube.

† Average power based on 0.1 duty.

Table 1—*Continued.* **Sperry Gyroscope Company—Reflex Klystrons**

Type	Frequency, megacycles	Resonator Voltage	Reflector Voltage (Negative)	Heater Voltage	Grid Voltage	Output	Tuning Range, megacycles	
							Mechanical	Electronic
3K27	750– 960	1,000	0–1,500	6.3	+20 to −200	1 w	210	10
3K23	950– 1,150	1,000	0–1,500	6.3	+20 to −200	1 w	200	10
2K41	2,660– 3,310	1,250	0– 750	6.3	+50 to −200	250 mw	650	17
2K42	3,300– 4,200	1,250	0– 750	6.3	+30 to −200	250 mw	900	15
2K43	4,200– 5,700	1,250	0– 750	6.3	+30 to −200	250 mw	1,500	15
2K44	5,700– 7,500	1,250	0– 750	6.3	+30 to −200	250 mw	1,800	15
2K39	7,500–10,300	1,250	0– 750	6.3	+30 to −200	250 mw	2,800	44

Table 1—*Continued.* **Sperry Gyroscope Company—2-cavity Oscillator/Amplifiers**

Type	Frequency, megacycles	Resonator Voltage	Grid Voltage	Heater Voltage	Output	Tuning Range, megacycles		Remarks
						Mechanical	Electronic	
3K21	2,300– 2,725	3,000	0—200	6.3	20w	425 Mc	10 Mc	
3K30/410R	2,700– 3,300	3,000	0—200	6.3	20w	600 Mc	10 Mc	10–14 db gain
3K22	3,300– 4,000	3,000	0—200	6.3	20w	700 Mc	10 Mc	

Table 1—*Continued.* **Sperry Gyroscope Special-purpose Tubes**

2K35	2,730– 3,330	3,000	0—200	6.3	25w	600 Mc	....	3-cavity, 2-stage cascade amplifier, 30 to 33 db gain
2K34	2,730– 3,330	3,000	0—200	6.3	16w	600 Mc	....	3-cavity oscillator buffer amplifier
2K47	(250– 280 2,250– 3,360)	1,000	0—200	6.3	125 mw	Input 30 Mc Output 110 Mc	....	2-cavity frequency multiplier
2K46	(2,730– 3,330 8,190–10,000)	1,500	0—200	6.3	10–70 mw	Input 600 Mc Output 1,810 Mc	....	3-cavity amplifier-frequency multiplier

GASEOUS CONDUCTION TUBES

By D. S. Peck

For definitions of gas tube, anode, cathode, etc., see pp. 4-03 to 4-06.

Arc. An arc is a discharge of electricity through a gas, characterized by a change in space potential in the immediate vicinity of the cathode which is approximately equal to the ionizing potential of the gas. (Proposed for Standards for Pool-cathode Mercury Arc Power Converters, AIEE.)

13. GASEOUS CONDUCTION

Gaseous discharges may be classified in two groups according to the mechanisms for producing ionization. The first group, "self-sustaining discharges," includes those in which the energy for maintaining the discharge is supplied directly by the discharge. The other group includes those that require some auxiliary power in addition to the energy of the discharge itself.

SELF-SUSTAINING DISCHARGES. One of the earliest known gaseous-discharge devices, the Crookes tube, is a typical example of a self-sustaining discharge. In that tube, cylindrical in form and containing a low gas pressure, a luminous discharge takes place if potential of sufficient value is applied between two electrodes. The appearance of the glow is shown diagrammatically with the corresponding voltage distribution in Fig. 1.

It will be observed that immediately adjacent to the cathode is the Crookes, or cathode, dark space. Across this space a large part of the total tube drop is concentrated. This drop is such that ions moving toward the cathode obtain sufficient energy to remove electrons from the cathode by bombardment and hence maintain ionization in the tube.

Following the cathode dark space is a luminous region, called the negative glow, in which some of the excited atoms are returning to normal, radiating energy in the form of light.

Following the negative glow is another dark space called the Faraday dark space, then another luminous portion called the positive column which extends to the anode.

This general type of discharge is used for the production of light in many neon and argon signs. It is also employed in protective tubes and in glow tubes used for voltage regulation, as well as in cold-cathode relay tubes.

Another form of self-sustaining discharge is found in tubes employing a pool cathode. Here the current densities run much higher than in a glow discharge, and the cathode dark space becomes very small so that extremely high gradients are present close to the emitting "spot." It is thus possible to release electrons without the high drop generally found in so-called glow discharge tubes similar to the Crookes tube. With a pool cathode, emission appears to be true field emission caused by voltage gradient, and not thermal emission caused by the high temperature of the spot.

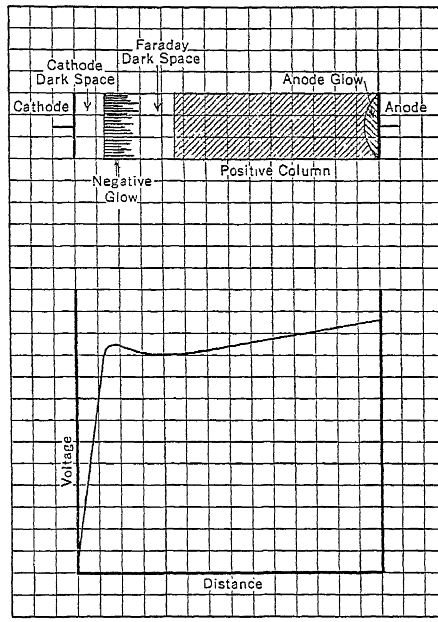


FIG. 1. Voltage and Glow Diistrbution in Glow Discharge

DISCHARGES REQUIRING AUXILIARY ENERGY. One of the commonest forms of this type of discharge is the hot-cathode tube. Here a heated filament or cathode, by virtue of the thermal energy imparted to the surface molecules, is able to release electrons. These electrons are drawn from the cathode by the anode field and after sufficient travel acquire enough energy to ionize the gas atoms by collision. The discharge is very similar in structure to the glow discharge from a cold cathode, except that the region of cathode fall is very much smaller and has a lower voltage drop since the electrons are actually ejected by the emitting properties of the cathode.

Other sources of ionization are possible, such as heat, photoelectrons, and radiation.

CONTROL OF THE DISCHARGE. If, in a simple tube with a cathode and an anode, a third element known as a grid is introduced between cathode and anode, it is possible to control the starting of the discharge. If the grid is made sufficiently negative with respect to the cathode, it produces a retarding field at the cathode in spite of the positive anode potential and most of the electrons emitted from the cathode are turned back without receiving sufficient energy to cause ionization. If the grid is then gradually made less negative, a critical voltage is reached where some electrons acquire enough velocity to ionize the gas. Within a few microseconds the ionization builds up until the current is limited only by the impedance of the external circuit.

Once ionization is complete, further changes in grid voltage have little effect on the discharge in most practical cases. If the grid is made negative, positive ions from the discharge will move toward it, causing a grid current to flow and blanketing the grid sufficiently to prevent its having any further effect on the arc. The grid becomes "sheathed" with positive ions, the thickness of the sheath depending upon the density of ionization and, to some extent, upon the grid voltage. The sheath thickness may be of the order of hundredths or thousandths of a centimeter under normal operating conditions. Only by using grids with very small openings and relatively large negative voltages is it possible to make the sheaths large enough to overlap and extinguish the discharge. If, however, the discharge ceases, as it would when alternating potential is used for the anode supply, the ionization diminishes to a low value or disappears in a relatively short period of time, and the grid may again exercise its control function.

VACUUM-TUBE CONTROL VS. GAS-DISCHARGE CONTROL. The characteristic of a grid-controlled gaseous-discharge tube of passing either no current or full current, together with the normal inability of the grid to stop the discharge, are two important differences from hard-vacuum grid-controlled tubes. Although both types possess the unidirectional conduction properties of a rectifier, the hard-vacuum tube with its continuous control may be likened to a rheostat in series with a circuit, whereas the gas-filled control tube may be considered similar to a switch which can be closed at any time but opened only when current is not flowing. Obviously, the hard-vacuum tube, which functions like a rheostat, must be capable of dissipating the power losses caused by pasage of current through the tube when the voltage across the tube is increased. The low voltage drop (from about 5 to 15 volts in most hot-cathode tubes) existing when the gas tube is passing current results in relatively low losses in the tube, and there is no plate dissipation whatever when the tube is turned off. Thus, a gas tube is capable of handling much heavier currents and more power than a hard-vacuum tube of the same size.

Because of the difference in control characteristic the circuit technique used with gas tubes is entirely different from that employed with hard-vacuum tubes. For instance, the gas tube may be used as a sensitive relay to operate a contactor when the grid voltage of the tube reaches a given value. With a voltage of adjustable phase relation on the grids it may be used as a controlled rectifier tube to control average output voltage. The gas tube may also be used in suitable circuits to change direct current to alternating current or alternating current of one frequency to alternating current of a different frequency.

14. THYRATRON TUBES

A thyratron is a hot-cathode, gas-discharge tube in which one or more electrodes are employed to control electrostatically the starting of the unidirectional current flow. (IRE Standard.)

These types of tubes cover an intermediate power range. Tubes are available in sizes up to 12.5 amp average current and up to 15,000 volts peak. (See Table 1.)

Similar gas tubes are built without grids for use as rectifiers. Since their major single application is in transmitting circuits, they are listed in article 13 with transmitter tubes.

CONSTRUCTION. Figure 2 shows a typical construction of a simple filamentary-type gas tube. The filament is in the form of a ribbon of nickel or nickel-cobalt alloy, formed in a helical or S-shaped form, and coated with an electron-emissive coating. A grid in the form of a cylinder is mounted as shown, and supported by a clamp from the lower stem. The grid may also be mounted on the stem leads and supported additionally from the glass at the top of the tube. The nickel anode in the form of a cup is mounted by means of another glass stem at the top. A washerlike cross piece in the grid determines the control characteristic by its relative spacing between cathode and anode and by the size and shape of a hole in the washer. The sides of the grid shield the control area from effects of charges collected on the glass walls and from extraneous fields. Carbonized nickel is frequently used for these parts because of its heat-radiating properties, but bright nickel is sometimes necessary in inert-gas-filled tubes. Argon, xenon, and mercury vapor are the most common mediums for gas tubes.

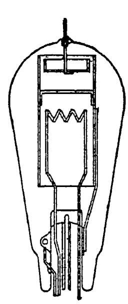


FIG. 2. Filamentary-type Thyratron Tube

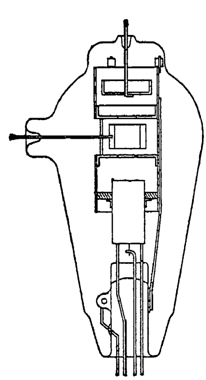


FIG. 3. Shield-grid Type of Thyratron Tube

FOUR-ELECTRODE CONSTRUCTION. Figure 3 shows a sketch of a shield-grid type of construction. The large shield grid, generally held at a fixed potential, permits the use of a small control grid of extremely high effective input resistance. The shield grid shields the control grid from the heat of the cathode and anode and minimizes the possibility of sputtered or evaporated material from the cathode contaminating the grid,

Table 1. Available Types of Thyratrons

Type	Designation	Anode Current			Peak Anode Voltage		Cathode *f* = filament *h* = heater		Ionizable Medium	Remarks
		Average	Peak	Averaging Time, seconds	Forward	Inverse	Voltage	Current		
2C4	S	0.005	0.020	30	450	450	2.5*h*	0.65	Gas	Negative control tube
297A	WE	0.010	0.060	....	250	250	1.75*f*	0.35	Gas	Negative control tube
546	GL-	0.020	0.100	15	500	500	6.3*h*	0.15	Gas	Shield grid tube
269A	WE	0.020	0.20	....	275	275	2.2*f*	0.55	Gas	Negative control tube
6D4	S	0.025	0.100	30	450	450	6.3*f*	0.25	Gas	Miniature negative control tube
233A	RX-	0.025	1.5	....	1,500	1,500	2.5*f*	2.5	Gas	Negative control tube
629	WL-	0.040	0.2	10	350	350	2.5*h*	2.6	Gas	Negative control tube
2051	Ray, Ch, GL-, RCA	0.075	0.375	30	350	700	6.3*h*	0.6	Gas	Shield grid tube
884	RX-, Ch, S, WL-, RCA, GL-	0.075	0.30	30	350	350	6.3*h*	0.6	Gas	Sweep-circuit tube
885	RX-, Ch, S, RCA, GL-, WL-	0.075	0.30	30	350	350	2.5*h*	1.5	Gas	Sweep-circuit tube
610	KU-	0.10	0.40	10	500	500	2.5*f*	6.5	Gas	Positive control tube
636	KU-	0.10	0.40	15	350	350	2.5*f*	7.5	Gas	Negative control tube
2D21	RCA	0.10	0.50	30	650	1,300	6.3*h*	0.6	Gas	Shield-grid tube
338A	WE	0.10	0.60	....	325	325	10.0*h*	0.50	Gas	Negative control tube
2050	Ray, GL-, Ch, WL-, RCA	0.10	1.00	30	650	1,300	6.3*h*	0.60	Gas	Shield-grid tube
502A	GL-, WL-	0.10	1.0	30	650	1,300	6.3*h*	0.60	Gas	Shield-grid tube
2A4G	S, Ch, Ray	0.10	1.25	45	200	200	2.5*f*	2.5	Gas	Negative control tube
178A	FG-	0.125	0.50	15	500	500	2.5*f*	2.25	Gas	Negative control tube
17	FG-, WL-	0.50	2.0	15	2,500	5,000	2.5*f*	5.0	Hg	Negative control tube
967	UE-	0.50	2.0	15	2,500	2,500	2.5*f*	5.0	Hg	Negative control tube
81A	FG-, WL-	0.50	2.0	15	500	500	2.5*f*	5.0	Gas	Negative control tube
98A	FG-	0.50	2.0	15	500	500	2.5*f*	5.0	Gas	Shield-grid tube
97	FG-	0.50	2.0	15	1,000	1,000	2.5*f*	5.0	Hg	Shield-grid tube
627	WL-, GL-	0.64	2.5	30	1,250	2,500	2.5*f*	6.0	Hg	Negative control tube
394A	WE	0.64	2.5	5	1,250	1,250	2.5*f*	3.25	Hg and gas	Negative control tube
3D22	RCA	0.75	6.0	30	650	1,300	6.3*h*	2.6	Gas	Shield-grid tube
C1B	EL-	1.0	8.0	....	450	700	2.5*f*	6.3	Gas	Negative control tube
303	CE-	1.0	8.0	....	450	700	2.5*f*	6.0	Gas	Negative control tube
302	CE-	1.5	4.5	....	1,000	1,000	2.5*f*	7.0	Hg	Negative control tube
287A	WE	1.5	6.0	5	500	500	2.5*f*	7.0	Hg	Negative control tube
323A	WE	1.5	6.0	5	500	500	2.5*f*	7.0	Hg and gas	Negative control tube
393A	GL-, WE	1.5	6.0	5	1,250	1,250	2.5*f*	7.0	Hg and gas	Negative control tube
3C23	GL-, WL-	1.5	6.0	5	1,250	1,250	2.5*f*	7.0	Hg and gas	Negative control tube
678	WL-, GL-	1.6	6.0	1 cycle	15,000	15,000	5.0*h*	7.5	Hg	Negative control tube
21	KY-	....	3.0	....	....	11,000	2.5*f*	10.0	Hg	Transmitter keying tube
628	KU-	2.0	8.0	30	1,250	2,500	5.0*f*	11.5	Hg	Negative control tube
305	CE-	2.0	12.0	....	850	1,700	2.5*f*	6.5	Gas	Negative control tube
27A	FG-	2.5	10.0	15	1,000	1,000	5.0*f*	4.5	Hg	Negative control tube
973	UE-	2.5	10.0	15	3,000	3,000	5.0*f*	6.75	Hg	Negative control tube
154	FG-	2.5	10.0	15	500	500	5.0*f*	7.0	Gas	Shield-grid tube
33	FG-, WL-	2.5	15.0	15	1,000	1,000	5.0*h*	4.5	Hg	Negative control tube
57	FG-, WL-	2.5	15.0	15	1,000	1,000	5.0*h*	4.5	Hg	Negative control tube
67	FG-	2.5	15.0	15	1,000	1,000	5.0*h*	4.5	Hg	Inverter tube
95	FG-	2.5	15.0	15	1,000	1,000	5.0*h*	4.5	Hg	Shield-grid tube
672	WL-, GL-	2.5	30.0	15	1,500	1,500	5.0*h*	6.0	Hg	Shield-grid tube
C3J	EL-	2.5	30.0	....	750	1,250	2.5*f*	9.0	Gas	Negative control tube
632-A	WL-	2.5	30.0	15	1,500	1,500	5.0*h*	6.0	Hg	Shield-grid tube
677	WL-	4.0	15.0	15	10,000	10,000	5.0*h*	10.0	Hg	Negative control tube
354A	WE	4.0	16.0	15	1,500	1,500	2.5*f*	16.0	Hg	Negative control tube
355A	WE	4.0	16.0	15	350	350	2.5*f*	16.0	Hg and gas	Negative control tube

Table 1. Available Types of Thyratrons—*Continued*

Type	Designation	Anode Current: Average	Anode Current: Peak	Anode Current: Averaging Time, seconds	Peak Anode Voltage: Forward	Peak Anode Voltage: Inverse	Cathode (f = filament, h = heater): Voltage	Cathode: Current	Ionizable Medium	Remarks
105	Ch, WL-, FG-	6.4	40.0	15	2,500	2,500	5.0h	10.0	Hg	Shield-grid tube
172	WL-, FG-	6.4	40.0	15	2,000	2,000	5.0h	10.0	Hg	Shield-grid tube
676	KU-	6.4	40.0	15	2,500	2,500	5.0h	10.0	Hg	Negative control tube
C6J	EL-	6.4	77.0	....	750	1,250	2.5f	21.0	Gas	Negative control tube
306	CE-	6.4	77.0	....	750	1,250	2.5f	18.0	Gas	Negative control tube
C6C	EL-	6.4	77.0	....	2,000	4,000	2.5f	24.0	Gas	Negative control tube
624	WL-	6.4	77.0	15	2,500	2,500	5.0h	10.0	Hg	Negative control tube
C16J	EL-	12	100.0	....	750	1,250	2.5f	31.0	Gas	Negative control tube
41	WL-, FG-	12.5	75.0	30	10,000	10,000	5.0h	20.0	Hg	Inverter tube
414	GL-, WL-	12.5	100.0	30	2,000	2,000	5.0h	20.0	Hg	Metal negative control tube

NOTES

Prefix	*Used by*
GL-	General Electric Company
FG-	General Electric Company
WL-	Westinghouse Electric Corporation
KU-	Westinghouse Electric Corporation
KY-	Eimac
UE-	United Electronics
EL-	Electrons, Inc.
CE-	Continental Electric Company
RX-	Raytheon

Letter	*Indicates*
S	Sylvania Electric Products, Inc.
Ch	Chatham Electronics
Ray	Raytheon
WE	Western Electric Company
RCA	Radio Corporation of America

with consequent reduction in grid emission. The shield grid may act also as an electrostatic shield to prevent sudden voltage fluctuations of the anode from inducing transient voltages on the grid with consequent loss of control. Of course, it is also possible to use tubes of this shield-grid type with "signal" voltages on both electrodes so that operation of the tube is a function of both grid potentials.

METAL TUBES. Metal envelopes are used for many industrial tubes, particularly in the larger sizes, because of the greater sturdiness and dependability of metal than of glass and to facilitate mounting the tubes on a panel. In the medium or smaller sizes this construction frequently has no real advantage over glass tubes because internal tube elements may be as subject to breakage as the envelope under conditions of shock or vibration.

15. VOLTAGE LIMITS OF THYRATRONS

PEAK INVERSE ANODE VOLTAGE. Peak inverse anode voltage is the maximum instantaneous anode voltage in the direction opposite to that in which the tube is designed to pass current. (IRE Standards on Electronics, 1938.)

The maximum peak inverse voltage which can be applied is a function of the shape and spacing of the electrodes, the current conducted, and the gas pressure. In thyratrons having argon or xenon or some other inert gas, the arcback potential is relatively independent of the tube temperature. In mercury-vapor tubes, however, the mercury-vapor pressure doubles roughly with every 10 deg cent increase so that the maximum peak inverse voltage is seriously affected. Figure 4 shows a typical curve of arcback voltage vs. temperature for a mercury-vapor tube. In rating a tube of this class, therefore, it is necessary to specify not only the peak inverse voltage but the maximum condensed-mercury temperature as well.

In designing a tube for high-voltage operation, it is sometimes necessary to constrict completely the space at the back, or top, of the anode, so that it is impossible for any discharge to take place between the anode and cathode around the outside of the grid.

PEAK FORWARD VOLTAGE. Peak forward voltage is the maximum instantaneous anode voltage in the direction in which the tube is designed to pass current. (IRE Standards on Electronics, 1938.) The same factors of tube geometry and current affect the maximum permissible peak forward voltage. In addition the grid must be so designed as to maintain the proper characteristics up to the voltage desired.

TESTS. The peak forward voltage may be tested by measuring the control characteristic at the maximum temperature ratings, although a common test is a full-load operation of the tubes in a circuit giving both peak inverse and peak forward voltages. This operation continues at the rated average current of the tubes for a length of time sufficient to stabilize the tube temperature, at which time the grid control is checked. The frequency of arcback is also observed. The severity of the test varies considerably with the current and voltage conditions at the time of current commutation from one tube to the next; therefore the type of operation circuit used is important. When simple circuits are used the voltage is frequently set above the rating to make the test sufficiently severe.

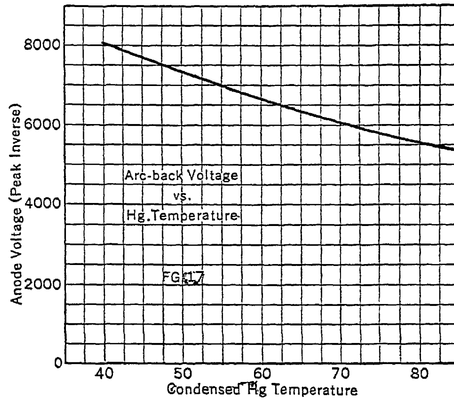


FIG. 4. Relation between Arc-back Voltage and Condensed Mercury Temperature in Typical Thyratron Tube

16. CURRENT LIMITS OF THYRATRONS

Most commercial tubes have cathodes made by coating a base metal such as nickel or a nickel alloy with one of two types of coatings. The first is barium oxide or a mixture of barium, strontium, and calcium oxides, formed by coating with the carbonates and reducing to the oxides by high temperature during the processing. The second is a molecular mixture of barium oxide and nickel oxide known as barium nickelate, formed by coating with barium carbonate and a nickel oxide before processing.

Cathode construction is of two general types, the filamentary and the indirectly heated. The filamentary type is simple in construction and has a relatively short heating time. The usual absence of heat shielding, however, reduces the efficiency, requiring more watts per ampere of emission, and increases the heat that the tube must dissipate. The indirectly heated type inherently lends itself to more efficient use of the cathode surface, thereby reducing the heating power required. It has the disadvantage of greater expense and longer heating time, a factor that is becoming more and more important.

By raising the cathode temperature it is usually possible to raise the emission per unit area and the emission per watt heating power. However, this raises the rate of coating evaporation and, if carried too far, shortens the life of the tube.

PEAK ANODE CURRENT. The peak current that a tube is capable of passing without harm depends upon the area, the temperature, and the geometry of the cathode surface, and somewhat upon the pressure of the gas. If the rated value is exceeded, the tube voltage drop may increase so that destructive ion bombardment of the cathode occurs. Also, particles of coating may be mechanically sputtered from the cathode to the grid, where they are a potential source of grid emission. There is the further possibility of sudden cessation of current flow, particularly at low gas pressure where the ion density is insufficient to neutralize the space charge at points where the discharge is the most dense. This is commonly known as "starvation" and may result in voltage surges in the circuit in which current has been flowing, sometimes breaking down insulation of the circuit components. The barium nickelate coating is less subject to this trouble since the coating is more conductive and will stand a higher arc drop without sputtering or surging.

AVERAGE ANODE CURRENT. Because of the essentially constant tube voltage drop over the current range, tube heating during operation is a function of average current, rather than rms current as in most electrical apparatus. Overloading causes excessive heating of all the tube electrodes and may result in failure because of (1) evolution of sufficient foreign gas to render the tube inoperative, (2) grid emission and loss of control due to high grid temperature, (3) arcback due to the high temperature of the anode, (4) short cathode life due to high evaporation, or (5) mechanical failure of the envelope or seals.

ANODE CURRENT AVERAGING TIME. When the load is fluctuating, as it might with a motor load or welding control, the average current must be calculated over a specified period of time chosen so as to include the worst current conditions. For example, consider two tubes having a maximum average current of 12.5 amp, a maximum peak current rating of 75 amp, and an averaging time of 30 sec in a biphase half-wave rectifier with sufficient inductance to make the d-c ripple negligible. These tubes could supply to the load 75 amp for 10 sec, 50 amp for 15 sec, or 25 amp for 30 sec out of every 30 sec without exceeding either peak or average rating.

GRID CURRENTS. In order to prevent overloading of the cathode or the grid through the grid circuit, a maximum instantaneous grid current and a maximum average grid current rating are generally given. Under most conditions of operation, grid currents are much less than these ratings.

ANODE SURGE CURRENT. The anode surge current is the current which would be conducted through the anode under fault conditions. The maximum surge current rating is a measure of the ability of the tube to withstand extremely high transient currents. The tube should carry the specified current for not longer than a given length of time in the event of short circuit, but it should not be expected to carry repeated short circuits without a reduction of life and the possibility of immediate failure. This rating forms a basis for set design to obtain best tube performance. If sufficient impedance is present to limit the fault current to this rating, not only will the tube be able to carry that current in the event of a fault in the circuit, but also the possibility of a fault in the tube itself seems to be reduced.

TESTS. Tests for maximum average current are generally made by operating the tube at full load and checking the other ratings which are dependent upon tube temperature, as described in other sections.

As has been stated, the peak current rating is a function of the emissive capabilities of the cathode, and the best test of this is, after all, satisfactory life while operating at that peak current. Several methods have been used for testing the emission, however. A simple test is to conduct an average current through the tube of a value between the average and the peak ratings and then measure the tube voltage drop with a d-c meter across the tube or with the deflecting plates of an oscilloscope connected from anode to cathode. This method does not discriminate very well between good tubes and bad except for some small tubes where the average current may be held fairly high for the test. It will suffice for a rough check on any tube, if used within the rating.

If, with the oscilloscope and a suitable d-c amplifier across the tube, the tube is allowed to conduct the peak rated current for only a few half-cycles each second, the cathode temperature is not altered appreciably by the load current, and the tube drop may be used as an accurate indication of cathode quality. Figure 5 shows a typical trace observed on a good tube, and Fig. 6 shows a low-emission tube in which the drop rises to an excessive value at the maximum current point.

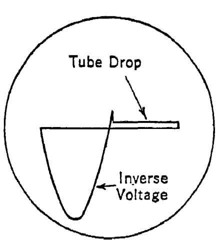


FIG. 5. Trace on Cathode-ray Oscilloscope Indicating Good Emission

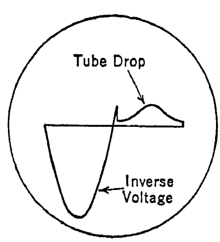


FIG. 6. Trace on Cathode-ray Oscilloscope Indicating Poor Emission

Another method of testing the emission is to conduct half-cycle pulses as above, but of increasing current magnitude, so that the point is finally reached where the cathode "sparks" or sputters, giving a broken voltage drop trace. The "sparking point" can be calibrated against the peak current rating to give a proper test.

In all such tests, the temperature of mercury-vapor tubes should be controlled quite accurately by an oil bath or controlled-flow air bath.

Surge current tests are generally made by the manufacturer on a given design of tube to insure a construction sufficiently rugged for ordinary service. Since these tests detract from the life of the tube, they are not part of the normal test procedure. In general, the test is made by passing the rated surge current through the tube with some protective device or control device arranged to open in a definite time, generally 0.1 sec. After the tube has been subjected to one or more overloads of this type, it is given an operation test for general performance.

17. CONTROL CHARACTERISTICS

Control Characteristic. The control characteristic of a gas tube is a relation, usually shown by a graph, between critical grid voltage and anode voltage. (IRE Standards on Electronics, 1938.)

Critical Grid Voltage. Critical grid voltage in a gas tube is the instantaneous value of grid voltage when the anode current starts to flow. (IRE Standards on Electronics, 1938.)

Critical Anode Voltage. The critical anode voltage of a gas tube is the instantaneous anode voltage when the anode current starts to flow.

As previously explained, once a thyratron is passing anode current the grid has little effect on the anode current, which is then limited only by the impedance of the load in the anode circuit. There are, therefore, no thyratron characteristic curves relating anode potential, grid potential, and anode current as there are for vacuum tubes. The relationship between anode voltage and grid voltage which just permits conduction, known as the control characteristic, is of considerable importance, however.

In an inert-gas-filled tube, there is a negligible change of characteristic with normal temperature changes. There are initial variations between tubes, however, and there are further changes in any given tube with life, the characteristic shifting slightly more negative as the emission becomes completely stable, and then shifting more positive as the end of the tube life approaches. Figure 7 shows the typical control characteristic of a small thyratron, the range of curves covering all variations between tubes and with life. This shows the equipment designer what range he may expect and must design for. In mercury-vapor tubes, the characteristic changes greatly with mercury temperature and the temperature must be specified for such a curve, or the range must be extended to include the variations within the published temperature limits. Figure 8 shows the variation in the control characteristic of a mercury tube with varying temperature.

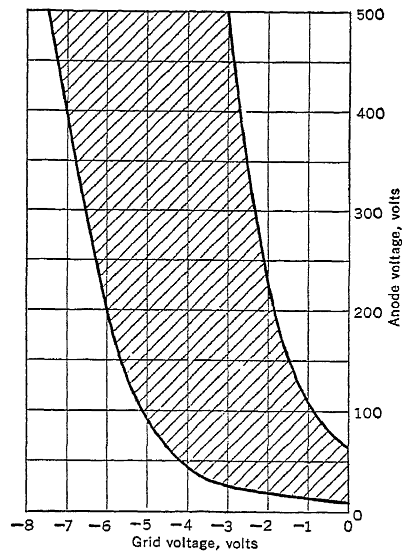


FIG. 7. Typical Thyratron Control Characteristic. (Shaded Area Shows Range of Characteristics.)

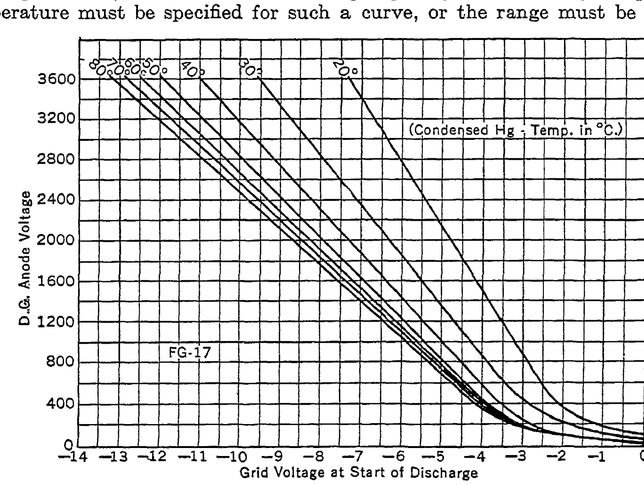


FIG. 8. Control Characteristics for a Mercury-vapor Thyratron

It will be noted that the critical grid voltages of the tubes shown are negative over the greater part of the operating range. As long as the grid is more negative than the cathode, electrons emitted from the cathode cannot flow to the grid, and there is no grid current from this source. Control is possible therefore with the use of a very small amount of grid power, which is the advantage of such a "negative grid" tube.

"Positive grid" tubes are available in which it is necessary to force the grid positive over the entire operating range to fire the tube. To assure non-conduction it is required only that the grid circuit be opened or the grid held at zero voltage or allowed to float. These application advantages are offset, however, by the fact that considerable power must be available to drive this type of tube, and sometimes a discharge takes place between grid and cathode before the anode-cathode path becomes ionized.

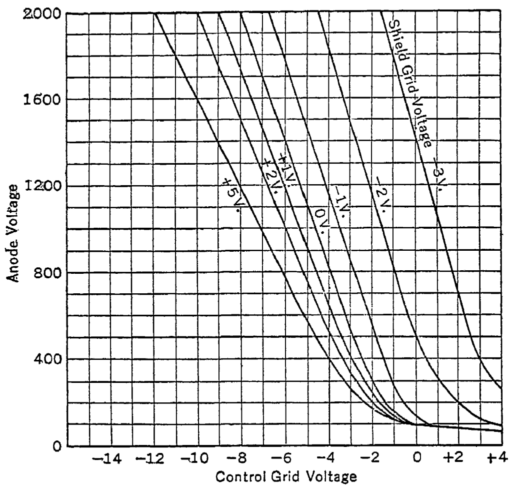


Fig. 9. Control Characteristics for Shield-grid Type of Thyratron, Showing Effect of Shield Potential

SHIELD-GRID CHARACTERISTICS. In four-electrode tubes the firing point becomes a function of the potential of both grids. Figure 9 shows how the control characteristic curve varies with shield-grid potentials. With negative shield-grid potentials, however, the variations between tubes become excessive with many types. The shield-grid signal may be used as a control to switch the characteristic in or out of the operating range of the circuit, and in some cases may even be used, with the necessary adjustments, to obtain a desired control curve.

GRID CURRENTS. The grid currents in a thyratron, both before and after discharge, are extremely important in determining the necessary grid power or the permissible impedance in the grid circuit.

Critical grid current in a gas tube is the instantaneous value of grid current when the anode current starts to flow. If a high grid resistance is used and this critical grid current

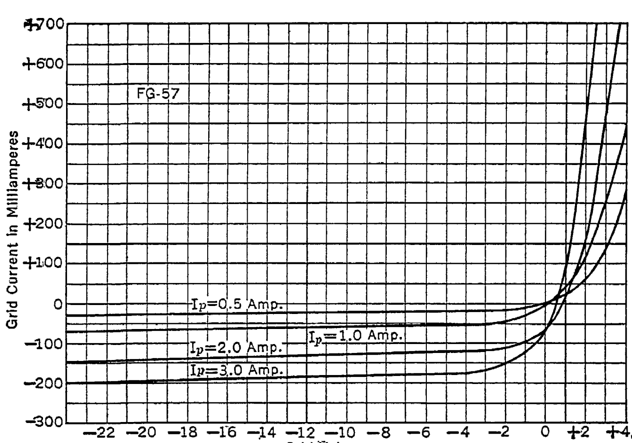


Fig. 10. Grid Voltage–Grid Current Curves after Breakdown in Three-electrode Thyratron

becomes excessive, owing to overheating and grid emission, enough voltage drop across the grid resistance may result so that the available grid supply voltage is not sufficient to control the tube. This current, then, becomes a limitation on the grid supply voltage and impedance and on the time constant with which the grid circuit can operate.

Figure 10 shows the grid current in a 2.5-amp average, three-electrode thyratron after discharge has occurred. Negative values indicate positive ion current to the grid from the discharge, and positive values indicate electron current. A four-electrode tube, otherwise similarly designed, would have roughly one-tenth the amount of grid current under these conditions.

Figure 11 shows the grid current for various anode voltages immediately before the start of discharge on the same three-electrode tube, and Fig. 12 shows the same for the equivalent four-electrode tube.

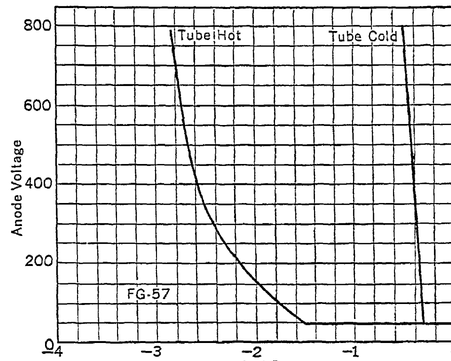


FIG. 11. Grid Current before Breakdown in Three-electrode Thyratron

TESTS. The control characteristic should be tested to the specified limits, usually at two or more values of anode voltage. A sufficient d-c control grid voltage should be applied to prevent conduction (or firing), through a low grid resistance. The specified d-c anode voltage is applied and the control grid gradually made more positive until breakdown occurs to the anode, at which time the critical grid voltage is observed. Condensed-mercury temperature should be controlled.

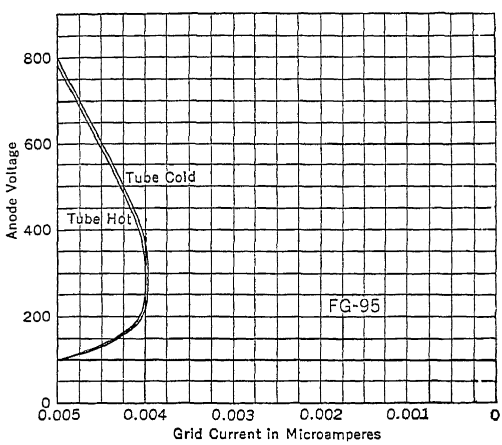


FIG. 12. Grid Current before Breakdown in Shield-grid Thyratron Tube

The critical anode voltage may be observed by holding the grid at zero voltage and increasing the anode supply until firing occurs.

The critical grid current is generally measured as shown in Fig. 13. The tube is first operated at the full average current rating as indicated by a d-c ammeter in the anode circuit, in order to heat all the parts to their operating temperatures. With the grid resistor r_g short-circuited, the grid supply voltage is then made more negative until the tube ceases to conduct. This is possible since an a-c supply is used and the tube does not conduct during the negative half-cycle. This voltage reading is denoted by V_1. Another reading is immediately taken with r_g in the circuit. With most tubes, a value of r_g between 10 and 100 megohms is sufficient to make the second reading, V_2, considerably higher than V_1. Since the actual critical grid voltage measured directly at the grid is the same in both measurements, the difference in the two readings must be accounted for as voltage drop in the resistor r_g. The grid current, therefore, is given as: $i = (V_2 - V_1)/r_g$. This reading of grid current includes currents from the ionized space to the grid, leakage currents, and grid emission. If the test is made without previous operation of the tube, the grid-emission factor will be eliminated.

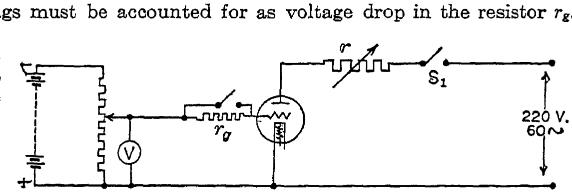


FIG. 13. Circuit Used to Check Grid Emission

DEIONIZATION TIME. Deionization time of a gas tube is the time required for the grid to gain control after interruption of the anode current. This varies with condensed-mercury temperature, anode current, anode voltage immediately after discharge, grid voltage, grid circuit impedance, and a number of other factors. Not only is the peak anode current just before discharge ceases important, but the wave shape is also, since in some cases the ionization due to an earlier peak current may decay less rapidly than the current, thus requiring a longer deionization time. Forcing the anode and grid negative immediately after the discharge has ceased decreases the deionization time since positive ions are then attracted to these electrodes, where they are neutralized.

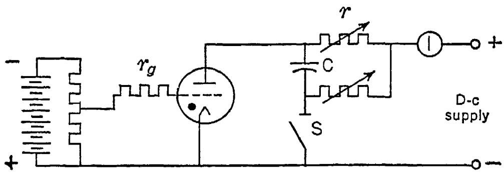


FIG. 14. Circuit Used to Check Deionization Time

To restore control to the grid it is not necessary to remove all the ions from the grid-anode space but only to reduce the number to a value sufficiently low so that the grid sheaths overlap enough to prevent discharge when a positive anode potential is again applied. Ions may still be present near the anode or near the cathode, with the grid having control. This shows why the grid voltage and stiffness of the grid circuit may be relatively more important in deionization than the anode circuit.

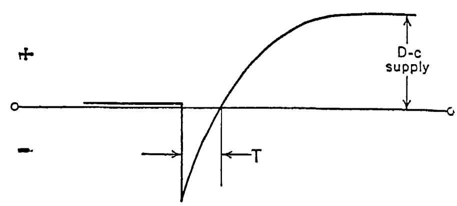


FIG. 15. Typical Anode-cathode Voltage Trace during Deionization Time

Since deionization time varies so widely with all these variables, it is essential to know all these conditions under which a measurement has been made. A recommended circuit for making such tests is shown in Fig. 14. The operation is briefly as follows:

The tube under test is conducting a specified d-c current. The capacitor C is connected in parallel with the load resistor R and therefore charges so that the negative capacitor plate is connected to the tube anode. When the switch S is closed, the capacitor makes the anode voltage of the tube instantly negative, thus stopping the discharge. The capacitor then recharges through the load resistor until the anode voltage of the tube and the capacitor voltage are equal to the supply voltage. The rate at which this anode voltage becomes positive depends upon the value of the capacitor C and of the load resistor, r. These values can easily be used to calculate the length of time from stopping of the discharge until the anode voltage again becomes positive, approximately the point at which the tube would conduct if deionization is not complete. At the capacitor setting at which the tube fails to control, the deionization time can be calculated from the formula $T = 0.693rC$. Any desired grid conditions may be used and the effect of these factors determined. Figure 15 shows a typical anode voltage trace during the test operation.

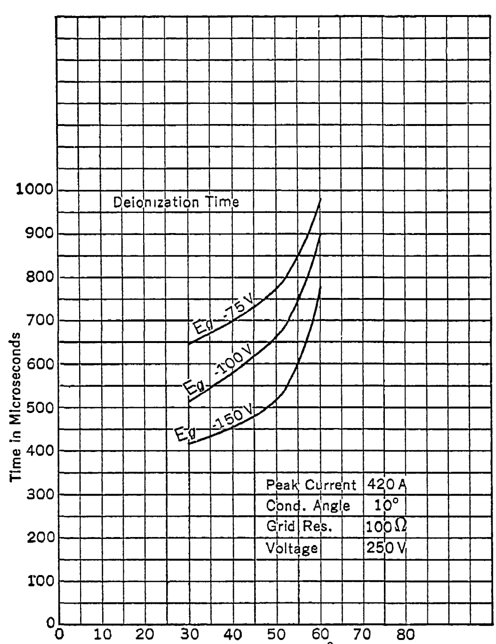


FIG. 16. Deionization Time Shown as a Function of Mercury Temperature

Other modifications of this circuit may be used to allow greater variation of anode voltages, or inverter circuits may be set up which will operate the tubes under more extreme conditions. Figures 16, 17, and 18 show the effect of some of the factors mentioned above.

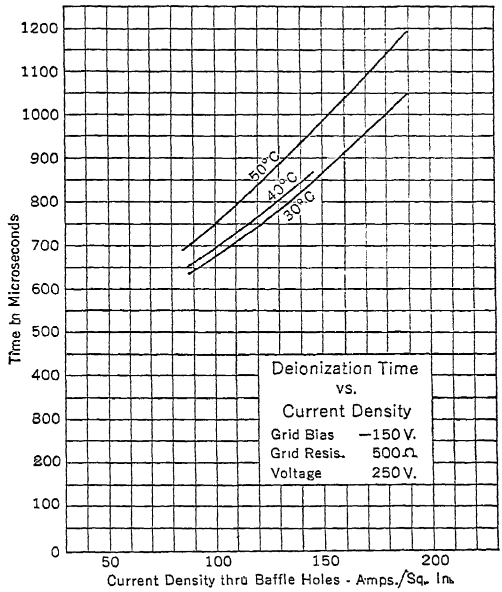


FIG. 17. Deionization Time Shown as a Function of Tube Current Density

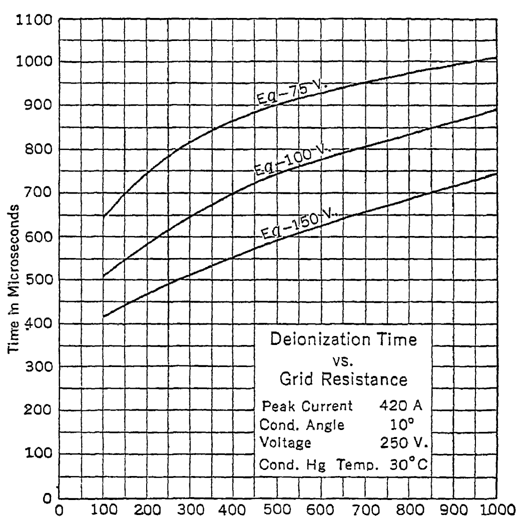


FIG. 18. Deionization Time Shown as a Function of Grid Resistance

18. PULSE THYRATRONS

As described in Section 9, Pulse Techniques, many circuits have been developed in which the currents are essentially square-wave pulses. One application of this sort provides modulated power for driving a magnetron; here a thyratron is used to discharge a pulse-forming line or network, previously charged to a high d-c voltage, through a pulse transformer into the load. The thyratron is in series with the line and load, and it carries a pulse current of magnitude determined by the voltage charge on the line and the combined impedances of the line and load.

PULSE VOLTAGES. During the charging of the line, the thyratron is subjected to the full instantaneous network voltage, so that it must be capable of holding off this voltage. It is generally desirable that this be done with zero grid bias on the tube, and consequently pulse tubes are designed so that they will fire only when considerable positive voltage is applied to the grid. This is generally of the order of 50 to 100 volts, although a higher voltage may be required by the manufacturer's data in order to insure that firing will occur with a minimum of time variation.

Time jitter is a common term for expressing the variation in the length of time from application of the grid firing voltage until start of the anode current pulse. Although the actual amount of this delay time is not so important, any variation tends to destroy the usefulness of systems in which the pulse may be as short as 0.1 microsecond. In order to keep the time jitter appreciably lower than this figure, the trigger pulse applied to the grid should have a steep wave front and a peak voltage appreciably higher than the minimum required to cause ionization. The internal impedance of the trigger supply must also be low enough to provide sufficient grid power to assist in accurate firing.

At the end of the anode-current pulse, depending on the ratio of load and line impedances, the voltage across the line, and hence across the tube, may become negative. The allowable negative voltage is limited because of the possibility of the remaining ionization causing arcback. After deionization is complete, the inverse voltage may be increased, generally to about the forward rating; deionization will generally occur in a fairly small percentage of the total time of application of inverse voltage and will allow the use of a-c charging voltages, other factors being satisfactory.

Tubes are tested for their voltage capacity by operation in a pulse circuit at the required forward voltage, peak pulse current, pulse width and repetition rate, and inverse voltage, and with a grid circuit which represents the minimum driving conditions which may be allowed.

PULSE CURRENTS. Pulse operation for thyratrons was not originally developed for industrial applications, and the life expectancies under such operation have been somewhat lower than those of other uses or other tubes. Cathode processing requirements are different, to the end of obtaining from 10 to 20 times the peak current density that is obtained from industrial tubes. These currents are limited in duration, however, to a pulse width of the order of 5 microseconds.

Many tubes designed for industrial operation are found to operate satisfactorily in low-power pulse circuits at much higher currents than their nominal ratings, and satisfactory pulse ratings have been developed for some industrial tubes.

Pulse tubes are generally rated for a maximum pulse repetition rate, pulse width, average current, and duty. For service in which none of these factors vary with time, the pulse width (in seconds) times the repetition rate (in pulses per second) will give the duty, or fraction of the total time, during which the tube is actually carrying a current pulse. This also serves as a relation between the peak and average currents, and thus limits the heating of the tube electrodes. In some applications it may be desirable to operate the tubes at a varying repetition rate; in this case the duty is sufficient to define the average current, and it becomes necessary to consider the minimum amount of time allowed between adjacent pulses to take care of deionization.

Because of the type of gas used in order to meet the strict deionization requirements, and the current densities used, the tube voltage drop during the pulse is inherently high. In addition, ionization requires some amount of time, during which time the arc voltage will be even higher. As a result, the tube voltage drop may rise to an initial value of 100 to 300 volts and then drop during the pulse to a value of 50 to 100 volts, these figures depending on the type of tube and other factors. A value may be read at some specific time during the pulse as an indication of tube quality, or the tube voltage may be plotted against the tube current on an oscilloscope, the shape of the resultant curve serving as a basis for judgment. In this case the voltage is generally higher during the rising current than during the falling current, and the area enclosed in the corresponding traces is an indication of the amount of ageing of the cathode during conduction. Information for the proper testing of any tube type may be requested of the manufacturer.

CONSTRUCTION. Although some industrial types of tubes have found pulse applications, these have been at low repetition rates and voltages, and most of the usable tubes have been designed specifically for pulse operation.

The cathodes are indirectly heated, since filamentary types have not been found satisfactory. Heating times are generally in the order of 1 to 3 min. This comparatively short heating time has been obtained by winding light heater wire either inside the cathode cylinder or outside a cylinder which is coated on the inside and which may have vanes projecting toward the inside to increase the emitting area. The alkali-earth oxides are used for the emissive coating.

The ionizing medium is usually hydrogen because of its light weight and high ion mobility and consequent speed of deionization. Especial care must be taken in processing, however, because of the susceptibility of hydrogen to cleaning up into the parts, particularly in the presence of contamination. For some of the tubes, high-purity alloys are used which are not required in the ordinary tube.

The anode must be completely shielded from any possible glow discharge or arc except the normal discharge, and it is surrounded at close spacing by the grid. In addition to the usual openings through the grid, an additional shield is placed directly below these openings and connected to the grid proper. This shield assures that the tube will maintain its highly positive characteristic and operate satisfactorily at high voltage.

Most of the applications have been met by three tube types, one operating at a pulse voltage of 3000 volts and a pulse current of 35 amp; the second at 8000 volts, 90 amp; and the third at 16,000 volts and 325 amp.

19. INSTALLATION AND OPERATION OF HOT-CATHODE THYRATRON TUBES

For any particular application complete data should be obtained from the tube manufacturer. Observation of the instructions will be amply repaid in satisfactory operation and long tube life.

FILAMENT CIRCUIT. The greatest single cause of unsatisfactory tube operation in the past has been incorrect filament voltage. Too low a filament voltage results in low emission and sputtering of the cathode which will give very short tube life and may prevent the tube from ever operating satisfactorily. Too high a filament voltage results in an excessive rate of evaporation of cathode material which results in short tube life. A filament-voltage variation of ± 5 per cent is generally permitted. Line-voltage variations outside the specified ± 5 per cent limits for a very short time, such as 5 or 10 sec, are not generally very harmful in the cathode-type tubes where the cathode temperature changes rather slowly. If excessive voltage variations are experienced and no corrective apparatus is available, it is sometimes possible to put the apparatus into operation by so adjusting the filament voltage that it does not drop more than 5 per cent below the rated value during the worst periods of sustained low voltage. Under these conditions the average filament voltage will probably be high and tube life will be shortened but not to the same extent as with too low a voltage.

CATHODE HEATING TIME. In some of the smaller quick-heating filamentary thyratrons it is possible, under some conditions of operation when the anode voltage and current are rather low, to apply both filament and plate voltages simultaneously without causing excessive damage. However, when full rated current is to be drawn or when a slower indirectly heated cathode is used, it is essential to provide some means whereby the tube will not pass anode current until the cathode is up to operating temperature. This may be accomplished by a switch in either the anode or load circuit or by proper bias on the grid. Wherever it can be economically justified, a time-delay relay or circuit may be used to give this protection. Such a relay will also provide protection in the event of a power failure.

TUBE HEATING TIME. An inert-gas-filled thyratron is ready to operate as soon as the cathode comes up to temperature. It should be noted, however, that in a mercury-vapor tube this is not always true. The condensed-mercury temperature must be within the rated limits before the tube is operated. Under conditions of low ambient temperature where it is desired to bring the tubes up to operating temperature quickly, tube enclosures and external heaters may be advantageous.

INITIAL HEATING. During shipment of mercury-vapor tubes the mercury may have become spattered over the tube electrodes. When first put into service the tube should be allowed to heat long enough to evaporate the mercury from the electrodes and to distribute it properly before anode voltage is applied. This may take a half hour or longer, depending upon the size of the tube.

ANODE CIRCUIT. Tubes should not be used in circuits where the voltages are higher than the rated peak inverse or peak forward voltages of the tubes. If there is a possibility that transients may be present, a cathode-ray oscilloscope or calibrated sphere gap across the tube may prove very useful in determining to just what voltages the tube is subjected.

CURRENT OVERLOADS. The low, essentially constant, voltage drop common to gas-filled tubes makes it possible to overload them injuriously to a greater extent than most hard-vacuum tubes, which have a rising voltage drop with increasing current. Once a thyratron tube is conducting, the current is limited only by the impedance of the load circuit. If it should be connected to a very low-reactance source of power with no load interposed, an abnormal current might flow, destroying the tube in a fraction of a second. A number of tubes are ruined in this way by experimenters, particularly those who are familiar with hard-vacuum tubes but who do not appreciate fully this important difference. For this reason *it is very important to make sure that some current-limiting impedance is in series with both the anode and grid leads when voltage is applied.*

This low, constant voltage drop makes it possible to impose heavy peak current overloads on the tubes without drawing excessive average current as indicated by a d-c ammeter directly in series with the tube. Unless sufficient reactance is present in the circuit, these peaks may be experienced when tubes are used in rectifier applications with capacitors connected directly across the output, when tubes are used to charge storage batteries, or when they are operating into any type of counter emf load. Although calculations may sometimes be made, it is generally safest to check tube currents with an oscilloscope when high peak currents are suspected.

REACTANCE OF POWER SOURCE. It is desirable to include enough reactance in the anode transformers or other sources of power to limit the tube short-circuit currents under any type of fault or failure to a value not greater than the surge current rating of the tube and to provide protective fuses or other interrupting devices to open the circuit in a reasonably short time, generally not in excess of 0.1 sec. Experience has shown that tubes operate much more satisfactorily in circuits in which this precaution is taken than in ones in which no such protection is provided. Various theories have been offered in explanation, but the fact has been well proved that high-reactance circuits give long tube life and low-reactance circuits give short tube life. In multitube circuits arcback of one tube often overloads the others, and again the circuit reactance must be depended upon to prevent permanent injury to these tubes. If excessively low-reactance transformers are used and if tube failures are encountered which show evidence of heavy overloads, such as blown-up stems or cracked seals, a small resistance or inductance placed directly in the anode lead of each tube will often stop the trouble.

GRID CIRCUIT. Usually the source of grid power itself has sufficient impedance to prevent excessive grid current from damaging the tube. If any doubt is felt on this score a grid resistor should be used to prevent accidents. However, in certain types of inverter circuits it is advisable to use a relatively stiff grid circuit to help speed up deionization. In the selection of a resistor the grid-current specification of the particular tube to be used should be considered so that difficulty will not be experienced through the loss of bias caused by the flow of grid current in the grid resistor. This resistor should, however, be large enough so that the voltage measured directly at the grid with the tube conducting does not exceed its rated value with the normal operating grid-supply voltage.

20. COLD-CATHODE TUBES

The term "cold-cathode tube" refers to tubes in which, in general, the discharge is the self-sustaining glow type previously described.

Since no cathode heating is required, the cathode life is not affected by standby operation, and these tubes are particularly adapted to relay circuits of infrequent operation, if the current requirement is sufficiently small, in addition to the familiar regulator and rectifier applications.

Cold Cathode. A cold cathode is a cathode operating at a temperature at which thermionic emission is negligible.

Starter (or control anode). A starter of a cold-cathode tube is an electrode ordinarily used in conjunction with a cathode to initiate conduction

Control Gap. A control gap of a cold-cathode tube is the conduction path between a starter and cathode in which conduction is ordinarily initiated.

Main Gap. A main gap of a cold-cathode tube is the conduction path between a cathode and a principal anode in which the principal conduction ordinarily takes place.

Anode Breakdown Voltage. The anode breakdown voltage of a cold-cathode tube is the anode voltage required to cause conduction to take place in the main gap when the control gap is not conducting.

Starter Breakdown Voltage. The starter breakdown voltage of a cold-cathode tube is the starter voltage required to cause conduction to take place in the control gap.

Transfer Current. The transfer current of a cold-cathode tube is the control-gap current required to cause conduction in the main gap with positive voltage applied to the anode.

Regulation. Regulation of a cold-cathode tube is the difference in voltage drop obtained over a range of conducted current.

Inverse Anode Current. The inverse anode (or starter) current is the electron current flowing from the associated circuit to the anode (or starter) of a cold-cathode tube.

Anode Drop. The anode drop of a cold-cathode tube is the main-gap voltage drop after conduction is established in this gap.

Starter Drop. The starter drop of a cold-cathode tube is the control-gap voltage drop after conduction is established in this gap.

VOLTAGE REGULATOR TUBES. Two-electrode, inert-gas-filled cold-cathode tubes are generally operated within a certain range of d-c load currents within which the tube voltage drop remains essentially constant. For instance, such a tube might break down or become ionized with a potential of 125 volts on the anode and then operate within 3 or 4 volts of a 90-volt tube drop within the specified current range. As with the hot-cathode tubes, the variation in supply voltage therefore appears across the load resistance, and it is essential that this resistance be sufficient to maintain the load current within the required value. Some calculations are generally necessary to insure that adequate starting voltage is also available when the supply voltage to be regulated falls to its minimum value.

Since regulator tubes are designed to conduct current in one direction only, they generally have one small electrode and one large electrode, the large one acting as the cathode. Tube life depends upon the current density on the cathode (and therefore the size of the cathode), since sputtering of the metal will occur if the element is overloaded.

THREE-ELECTRODE COLD-CATHODE TUBES. A third electrode or starter may be introduced into a cold-cathode tube to control the starting of the discharge. As with hot-cathode tubes the starter has no appreciable control once the discharge has occurred in the main gap until the anode potential has dropped below the tube voltage drop (sometimes called sustaining voltage).

The use of these tubes as relays is obvious because of their control characteristics. Another property of such a tube is that, although some inverse anode current will be conducted with negative anode voltage, the breakdown voltage in this direction is somewhat high and the anode drop increases rapidly as current increases. Figure 19 shows this characteristic. With the starter connected to the anode through a high resistance, the forward breakdown voltage will be much below the normal main-gap breakdown value and the tube will operate satisfactorily as a rectifier. The circuit must be such, of course, that peak inverse anode current or voltage ratings are not exceeded.

The starter drop is quite independent of control-gap current, and the tubes are used as voltage regulators with the cathode and starter as electrodes.

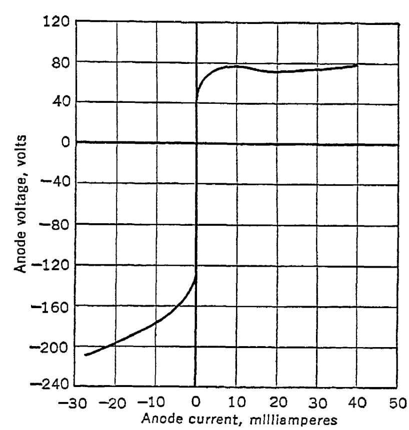


Fig. 19. Typical Cold-cathode-tube Characteristic

The mechanism of operation of the starter is similar to that of the grid in a positive-grid thyratron. When positive voltage is applied to the starter, a small electron current known as a Townsend current will flow from the cathode. As the voltage and the Townsend current are increased, a point is reached at which ionization occurs in the control gap. The control-gap current may then be increased by the control circuit until the transfer current is reached, at which point (depending upon the anode voltage) conduction occurs to the anode. At the anode breakdown voltage, conduction will occur without ionization in the control gap, and therefore the transfer current is zero. At the other extreme, the tube will not sustain a discharge at less than the anode drop, and hence the transfer current required at this point is infinite. Figure 20 shows a typical transfer-current characteristic. It is apparent that the anode operating voltage, when starter control is expected, must lie somewhere between the anode drop and the anode breakdown voltage.

The control characteristics are subject to variation during life but may be expected to remain constant within 5 or 6 volts during most of life. Variation during shelf life may be of the same order of magnitude depending upon the length of storage and the light conditions for those tubes which do not have an opaque coating. The effect of light is negligible for medium light levels but will be appreciable at levels approaching darkness or direct sunlight, the numerical value of breakdown voltages decreasing as light intensity is increased, and conversely with decreased intensity. Most variations occurring during storage will disappear after a few seconds of operation.

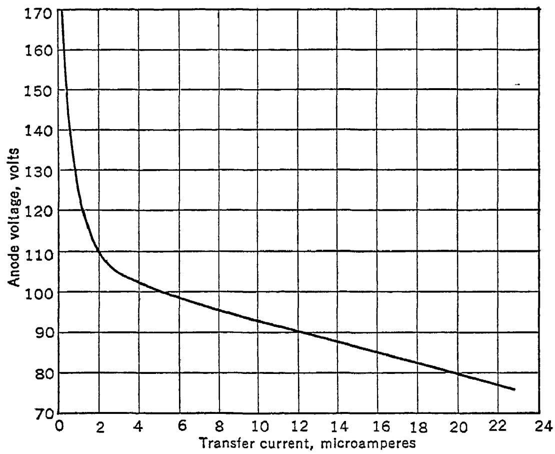


FIG. 20. Typical Transfer Current Characteristic

TEST FOR COLD-CATHODE TUBES. Because of the effect of light and storage upon control characteristics, it is generally advisable that all tests be made with moderate illumination and with a few seconds of current conduction immediately before the test in order to stabilize the readings. Because of the low values of transfer current some errors may be introduced if capacitance effects are ignored, and it is therefore important that the starter resistance be placed immediately adjacent to the starter electrode.

Breakdown voltage is measured by applying a positive d-c voltage to the anode and increasing it until the tube conducts current. The minimum value of voltage required to start conduction is measured. After the tube conducts, the tube drop may be measured at specified current values and the regulation thereby determined. For three-electrode tubes the anode breakdown voltage is measured as above, except with the starter connected to the anode through a resistance generally not exceeding 50,000 ohms.

Transfer current is measured by applying the specified d-c anode potential and a positive starter voltage, with sufficient starter resistance to limit the starter current to less than the transfer value. The starter current is then increased until conduction takes place in the main gap, at which point the transfer current is measured by means of a microammeter in series with the starter. Care must be exercised to insure that the starter is electrically connected at every instant during this test.

STROBOTRON TUBES. A modification of the three-electrode cold-cathode tube described is commercially called the Strobotron. This tube commonly has two control electrodes or grids together with the cathode and anode. The cathode is coated with a cesium compound, which breaks down during conduction and liberates free cesium. The tube is inert-gas-filled, and conduction occurs by formation of a cathode spot on the metallic cathode surface by concentration of a glow discharge. The voltage drop is therefore much lower than that of a glow discharge. These tubes are commonly used with neon gas as a source of high-intensity light for stroboscopic purposes but may also be used for control or relay purposes. Capacity is limited by cathode heating during conduction and by ion bombardment of the cathode as with other cold-cathode tubes. For this reason, although very high peak currents can be handled, the average current rating is generally quite low.

A capacitance is generally used across the tube in operation. This capacitance is discharged through the tube in order to provide very high instantaneous current. It also serves as a means of stopping the discharge, since, as the capacitance discharges to the tube drop, the conduction changes to a glow rather than an arc. This immediately raises the tube drop and the conduction ceases, the tube becoming deionized before the capacitor recharges to sufficient voltage to maintain even a glow discharge. Operation at too high a frequency, or with improper circuit conditions, will frequently maintain a glow discharge in the tube and lose the advantage of the tube's capacity for high peak currents.

With this tube as with other cold-cathode tubes life is a function of average current and conduction time, there being no appreciable deterioration during storage.

Available types of tubes are shown in Table 2.

Table 2. Available Types of Cold-cathode Tubes

Type	Desig-nation	Cathode Current, milliamperes		Anode Voltage Drop	Breakdown Voltage s = starter a = anode	Remarks
		Average	Peak			
874	RCA	10–50		90	125a	Voltage regulator
991	RCA	2	3	48–67	87a	Voltage regulator
BR	Ray	50		60		Rectifier
OA2	S, Hy, RCA	5–30		150	155a	Miniature voltage regulator
OB2	Hy	5–30		105	133a	Miniature voltage regulator
OA3/VR75	GL-, S, RCA	5–40		75	100a	Voltage regulator
OB3/VR90	GL-, S	10–30		90	110a	Voltage regulator
OC3/VR105	GL-, S, Hy, RCA	5–40		105	115a	Voltage regulator
OD3/VR150	GL-, S, RCA, Hy	5–40		150	160a	Voltage regulator
1B46	S	1– 2		79–85	225a	Miniature voltage regulator
1B47	S	1– 2		75–90	225a	Miniature voltage regulator
1B64	S	1– 2		65–75	225a	Miniature voltage regulator
313C	WE	35	100	75 at 20 ma	70s	Control tube
313CC	WE	18	50	75 at 20 ma	72s	Control tube
359A	WE	15	40	75 at 10 ma	75s	Control tube
395A	WE	13	35	75 at 10 ma	77s	Control tube
OA4G	S, RCA	25	100	70	75–90s	Control tube
1C21	RCA	25	100	73	66–80s	Control tube
618	KU-	15	100	180		Control tube
1B48	Ray	6	50	100 at 6 ma	800a	High-voltage rectifier

Prefix	*Used by*
GL-	General Electric Company
KU-	Westinghouse Electric Corporation

Letter	*Indicates*
WE	Western Electric Company
RCA	Radio Corporation of America
Ray	Raytheon
S	Sylvania Electric Products, Inc.
Hy	Hytron Radio and Electronics Corporation

21. POOL-CATHODE TUBES

Pool Cathode. A pool cathode is a cathode in which the principal source of electron emission is a cathode spot on a metallic pool electrode.

Cathode Spot. A cathode spot is an area on the cathode of an arc from which electron emission takes place at a current density of thousands of amperes per square centimeter and where the temperature of the electrode is too low to account for such currents by thermionic emission.

Pool Tube. A pool tube is a gas tube with a pool cathode.

Single-anode Tube. A single-anode tube is an electron tube having a single main anode.

Multianode Tube. A multianode tube is an electron tube having two or more main anodes and a single cathode.

Ignitron. An ignitron is a single-anode pool tube in which an ignitor is employed to initiate the cathode spot for each conducting period.

Excitron. An excitron is a single-anode pool tube provided with means for maintaining a continuous cathode spot.

Pumped Rectifier. A pumped rectifier is a rectifier which is continuously connected to evacuating equipment during operation.

Sealed Tube. A sealed tube is a tube which is hermetically sealed after degassing.

Ignitor. An ignitor is a stationary electrode which is partially immersed in the cathode pool and has the function of initiating a cathode spot.

As mentioned previously, tubes of the pool-cathode type make use of the self-sustaining form of discharge. The cathode dark space is very small because of the high current density, and the resulting high voltage gradient is assumed to cause field emission and thus maintain the high current density at an arc voltage drop near that of the ionization voltage.

Since there is no emissive material to be damaged, the pool cathode is capable of carrying extremely high currents without any deterioration. Tubes having such cathodes can therefore be used in applications where very high peak currents are required for short periods of time, or where it is desirable that a high short-circuit or arcback current be allowed without damage to the tube. This current can be much higher proportionally than for a hot-cathode tube. The pool cathode also has the advantage of not requiring heating time.

On the other hand, the pool tube requires auxiliary circuits for maintaining ionization or for starting the ionization during each operating cycle. The size of the tube and the corresponding volume of ionization increases the difficulty of building high-voltage tubes, or those with rapid deionization. Further, because of the amount of current capacity, water cooling is required for most types, and corresponding protection must be provided to insure proper water flow and water temperature at all times.

Available types of pool tubes are listed in Table 3.

Table 3. Available Pool-cathode Tubes

Type	Designation	Typical Ratings
415	GL-	300-kva, 22.4-amp average, welder control ignitron (sealed)
681/686	WL-	300-kva, 22.4-amp average, welder control ignitron (sealed)
652/657	WL-	600-kva, 56-amp average, welder control ignitron (sealed)
271	GL-	600-kva, 56-amp average, welder control ignitron (sealed)
235-A	GL-	1200-kva, 140-amp average, welder control ignitron (sealed)
651/656	WL-	1200-kva, 140-amp average, welder control ignitron (sealed)
655/658	WL-	2400-kva, 355-amp average, welder control ignitron (sealed)
258-A	GL-	2400-kva, 355-amp average, welder control ignitron (sealed)
259-B	GL-	600 volts d-c, 200 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c, 150 kw
679	WL-	600 volts d-c, 200 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c, 150 kw
653-B	WL-	600 volts d-c, 500 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c 300 kw
238-B	GL-	600 volts d-c, 500 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c, 300 kw
ES-8-01	Allis-Chalmers	600 volts d-c, 500 kw (output of 6 tubes),* rectifier excitron (sealed) 250 volts d-c, 300 kw
507	GL-	600 volts d-c, 1000 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c, 600 kw
688	WL-	600 volts d-c, 1000 kw (output of 6 tubes),* rectifier ignitron (sealed) 250 volts d-c, 600 kw
427	GL-	350-volt peak 10-amp glass demonstration ignitron (sealed)
506	GL-	9000-volts d-c, 8100 kw (output of 6 tubes),* rectifier and inverter ignitron (sealed)
8-in. tank	Westinghouse	600-volts d-c, 1000 kw (output of 6 tanks),* rectifier ignitron (pumped)
10-in. tank	General Electric	600-volts d-c, 1000 kw (output of 6 tanks),* rectifier ignitron (pumped) 250-volts d-c, 625 kw
10-in. tank	General Electric	3000 volts d-c, 1500 kw (output of 6 tanks),* rectifier ignitron (pumped)
12-in. tank	Westinghouse	600 volts d-c, 1500 kw (output of 6 tanks),* rectifier ignitron (pumped) 250 volts d-c, 750 kw
6-EP-20-01	Allis-Chalmers	600 volts d-c, 1500 kw (output of 6 tanks),* rectifier excitron (pumped) 250 volts d-c, 750 kw
16-in. tank	General Electric	600 volts d-c, 2000 kw (output of 6 tanks),* rectifier ignitron (pumped) 250 volts d-c, 1000 kw
16-in. tank	General Electric	3000 volts d-c, 4000 kw (output of 6 tanks),* rectifier ignitron (pumped)
15-in. tank	Westinghouse	600 volts d-c, 2000 kw (output of 6 tanks),* rectifier ignitron (pumped) 250 volts d-c, 1000 kw
20-in. tank	General Electric	600 volts d-c, 3000 kw (output of 6 tanks),* rectifier ignitron (pumped) 250 volts d-c, 1500 kw
22-in. tank	Westinghouse	600 volts d-c, 3000 kw (output of 6 tanks),* rectifier ignitron (pumped) 250 volts d-c, 1500 kw
6-EP-20-11	Allis-Chalmers	1750 volts d-c, 2500 kw (output of 6 tanks),* rectifier and inverter excitron (pumped)
HF-26	Allis-Chalmers	1000 cycles, 300 kw, multianode frequency changer (pumped)

* These are typical operating conditions rather than absolute maximum ratings.

WL-, prefix used by Westinghouse Electric Corporation; GL-, prefix used by General Electric Company.

CLASSIFICATION. Pool-cathode tubes may be classified as sealed or pumped units. In general, the single-anode sealed units are of comparatively lower current capacity, now being built in the range from about 50 amp average to about 400 amp average. After the initial exhaust treatment these tubes are permanently sealed and no further pumping is required. Pumped units have been built largely of metal in either single-anode units or in a multianode form commonly called a tank. These units have a pump or system of pumps operating to maintain vacuum. Current capacity of these tanks is in the order of 1000 amp average. Pool tubes may be classified still further as tubes in which there is continuous ionization and those in which the arc is initiated each time conduction starts and is extinguished at the end of each conducting period.

MERCURY-ARC RECTIFIERS. The mercury-arc rectifier is a familiar example of a pool tube in which continuous ionization is maintained.

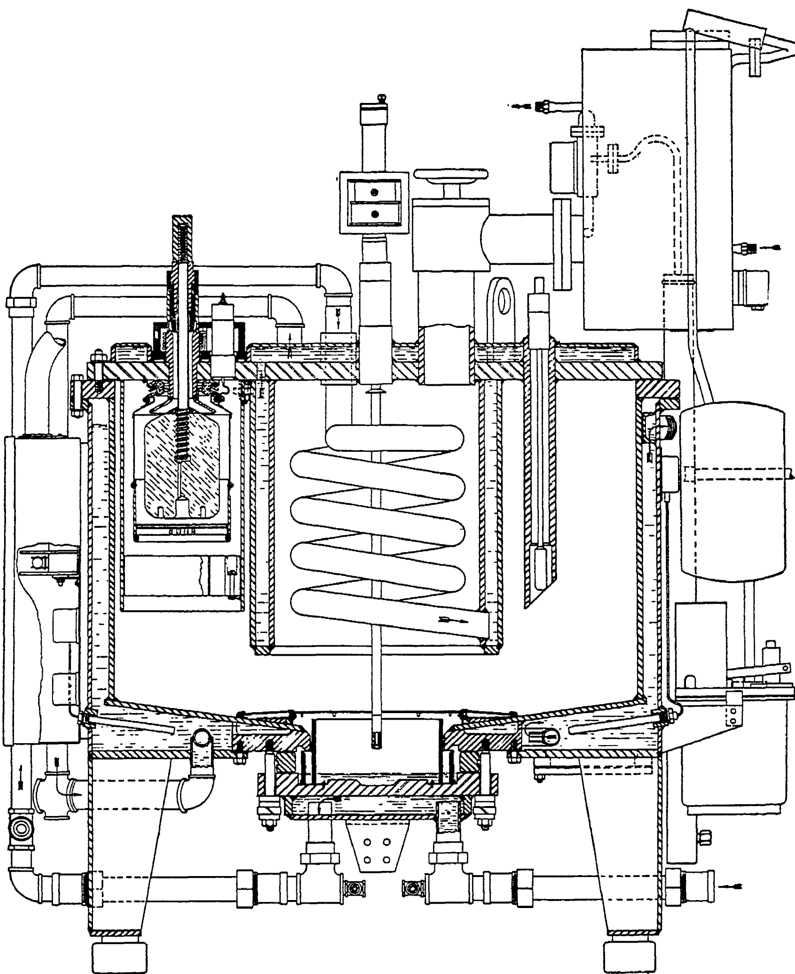


Fig. 21. Typical Pumped Mercury-arc Rectifier Tank. (Courtesy of G. E. Co.)

These types are made in both glass and metal. The glass types are sealed units with the mercury pool in the bottom and with glass arms extending out from the main body of the tube, slightly above the pool level. An anode is sealed into the end of each arm, which may or may not be bent, depending on the inverse voltage rating of the tube. Some two-anode types are made for lighting and railway service, but the more extensive applications are in three- or six-anode designs. These applications are mainly outside the United States and are in railway service and other power usage. A typical six-anode tube would have a capacity of the order of 400 amp average at 600 volts.

The arc spot of such tubes is generally formed by tilting the tube so that the mercury pool forms a circuit with a starting electrode. When this circuit is broken an arc is initiated and is maintained by one or more auxiliary anodes. These anodes form with the pool a low-voltage rectifier the purpose of which is to keep enough current (6 or 7 amp) flowing through the tube to maintain the cathode spot. As the tube current increases, the cathode spot breaks up into several spots each carrying about the same amount of current.

These tubes may be given grid-control characteristics by the addition of a grid around the anode in each arm.

Figure 21 shows an example of a metal multianode pumped rectifier tank rated for 1000 kw at 600 volts. Such tanks operate like the glass tubes except that the arc is initiated by a probe electrode in the center of the pool. This electrode is magnetically controlled and makes and breaks contact with the pool in order to form the arc spot. Ionization is then maintained by auxiliary anodes such as the one shown on the right-hand side of the tank, extending from the top cover into the main volume. One of the main anodes is shown at the left, with a suitable grid structure surrounding and with additional baffling below. The bottom surface of the tank slopes to allow the return of condensed mercury to the pool in the center, and a water-cooling system extends completely around the outside of the tank, under the mercury pool, and through a cooling coil and cylinder in the center of the tank. This provides a maximum of cooling area in contact with the arc and thus provides the best control of operation. Direction of water flow is indicated by arrows in the cooling system. The exhaust connection, vacuum controls, and pumps are shown at the top and right of the tank.

Such tanks are built with voltage ratings as high as 3000 volts direct current and with currents of the order of 1000 amp.

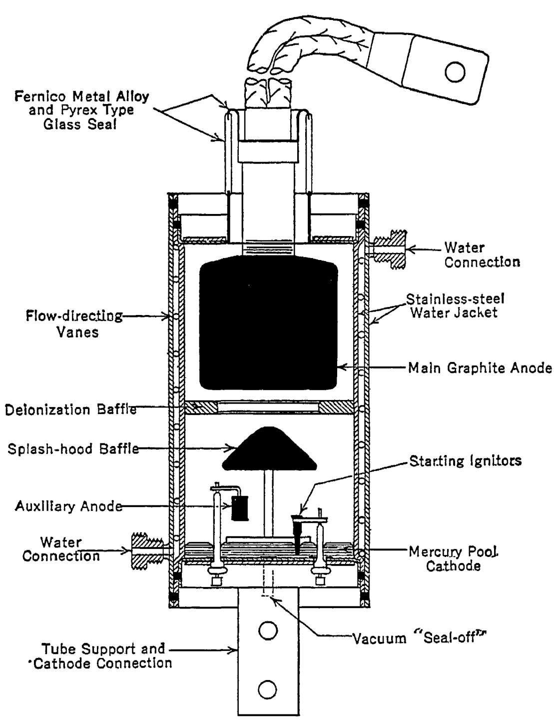


FIG. 22. Sealed Ignitron for Power Rectifier Service. (Courtesy of G. E. Co.)

IGNITRONS. Figure 22 shows a typical sealed ignitron tube. The two enclosing cylinders provide a path for cooling water, and, since the temperature of the inner tube wall is important in determining mercury-vapor pressure and operating capacity of the tube, it is essential that the required water flow be provided during operation. The graphite anode is designed for a maximum transfer of heat from its face and for a minimum of heat generation from the high current flow. Mercury is thrown up from the pool by the action of the cathode spots, and a splash-hood baffle is placed directly above the center of the cathode to prevent splashes of mercury from hitting the anode or upper tube walls and thereby increasing the possibility of arcback. The purpose of a deionization baffle is easily seen, since it reduces the distance from the ionized area to a metal part. This part is particularly important in tubes designed for rectifier service, where the deionization requirements are more severe than in welder use. The auxiliary anode, also, is used for rectifier service, where it is desirable to maintain ionization over the conducting period in order to provide for main anode currents below the value at which the arc becomes un-

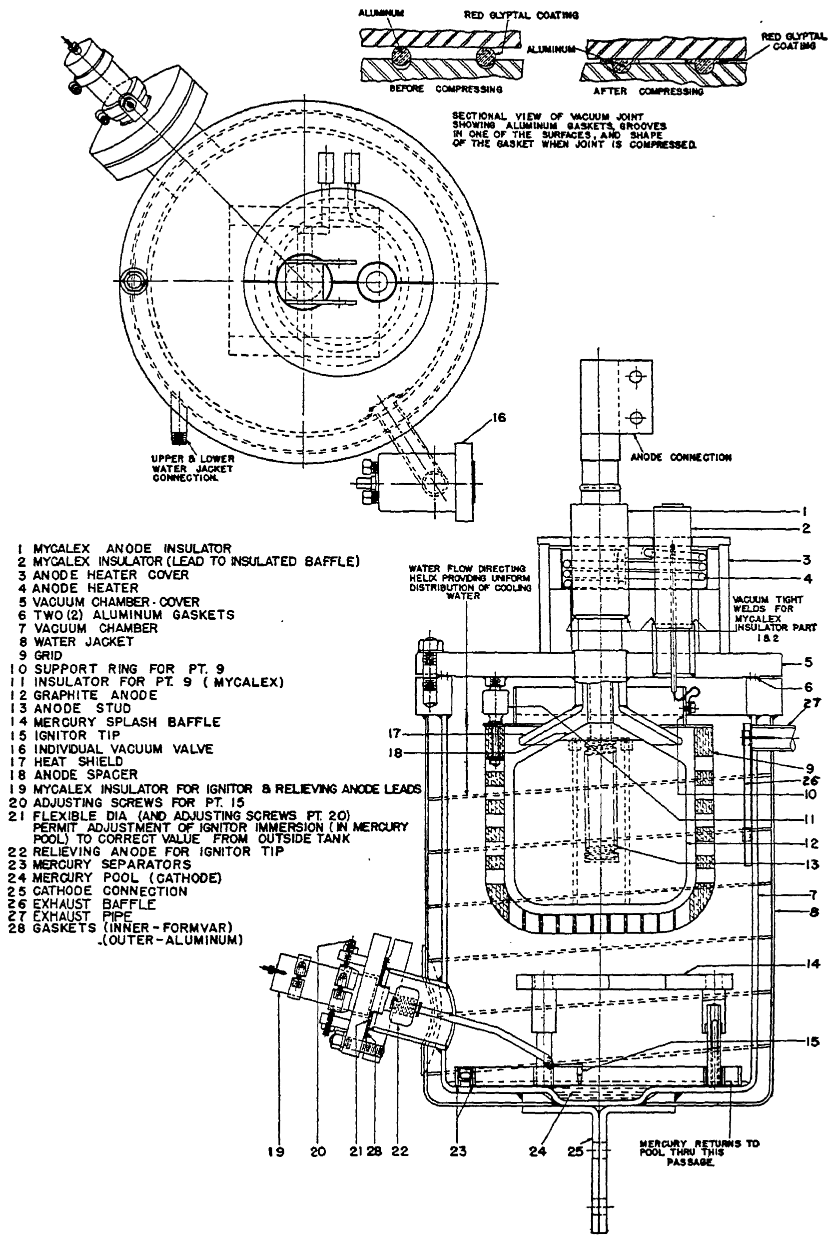


FIG. 23. Pumped Ignitron Tank. (Courtesy of G. E. Co.)

stable, usually 3 to 4 amp. Tubes for welder applications may be built without the two baffles and the auxiliary anode. On the other hand, tubes built for high-voltage rectifier use may have several baffles or grids surrounding the anode in order to provide high-voltage control and to reduce deionization time, and they may also have two or more ignitors in order to assure operation in the event of an ignitor failure.

Figure 23 shows a typical pumped ignitron tank, with vacuum-tight joints and flexible ignitor support to allow adjustment of the ignitor position; the other parts are similar in use to those in the sealed tube. The sealed and pumped units are used in similar applications.

IGNITOR CHARACTERISTICS. The ignitor as shown in Fig. 22 is shaped somewhat in the form of a pencil and is composed largely of boron carbide. The composition is such that the high temperature of the arc has practically no effect upon the ignitor surface, and there is no tendency for the mercury to wet the surface unless foreign material becomes deposited at the contact area.

The ignitor acts simply as a resistor; a current flows through it into the cathode pool when voltage is applied between the ignitor terminal and the cathode. When sufficient current flows, an arc is started between ignitor and pool, and conduction is established through the tube provided that positive anode voltage is applied. Resistance of the ignitor may vary from 10 to 100 ohms, and ignition may occur at instantaneous currents of 1 or 2 amp with 100 to 200 volts or at 15 to 30 amp with but a few volts applied. The manufacturer's data should be consulted to determine the exact ignitor requirements to insure proper operation throughout the life of the tube.

Formation of the arc may be caused by one or both of two mechanisms. At the point where the ignitor is immersed the mercury forms a meniscus, so that its surface is separated from the ignitor surface by very small distances at some points. High-voltage gradients exist across this space and cause a discharge to start. Another possible method of ignitor operation is that of developing heat at the several contact points of the rough ignitor surface with the mercury surface. The rapid heating at these sharp contacts due to the ignitor current flow causes the mercury to vaporize rapidly and break contact, thereby starting the desired arc.

Conduction to the anode occurs in the order of 20 to 200 microseconds after application of the critical ignitor voltage or current.

Ignitor ratings include the maximum allowable peak voltage and peak current ratings, which insure that the ignitor will not be eroded or otherwise affected during operation. A maximum average current rating limits the heating which may contribute to short life. The maximum required peak voltage and current specify the worst conditions of current or voltage expected for consistent firing, and a circuit designed to provide at least these requirements will provide satisfactory operation during the tube life.

Since the arcback characteristic of ignitrons depends upon current, voltage and temperature, the tubes are rated for more than one load current depending upon the operating voltage and duty cycle. For tubes in rectifier service these ratings take the form of two or more discrete current ratings at corresponding voltages or length of time of operation. For welding tubes, a curve of demand kva (rms current times rms circuit voltage) vs. average tube current is used, thus taking into account the on and off times of the welder as well as the maximum current carried.

IGNITRON TESTS. Because of the size of the tubes and equipment involved ignitrons are tested where possible in an operation circuit which closely approximates the application conditions. Tests at the maximum current, voltage, and duty cycle should give an indication, within a very short time, of the quality of the tube. The arcback rate is frequently used as a determining factor of this quality. The water-cooling coils are tested for their capacity to withstand expected water pressures, but more important than this test may be the user's test of the purity of the water being used, since sediment or scale will seriously reduce heat transfer from the inner wall of the tube.

Ignitors are generally tested in a simple manner by measuring the resistance from ignitor terminal to cathode pool under conditions of normal mercury level. A much more severe test is to apply a specified firing voltage and measure the time required for the tube to conduct current, or to measure both the ignitor current and ignitor voltage at which conduction takes place.

EXCITRONS. The excitron is a currently available type of metal rectifier, built in both sealed and pumped units. This type has the advantages of other single-anode rectifiers for production and maintenance and has the firing characteristics of the multianode tanks in which ionization is maintained continuously.

In general construction, commercially available excitrons are similar to the sealed or pumped ignitron except that the cathode is insulated from the tube wall, as it is in the metal mercury-arc rectifiers. This prevents travel of the arc to the wall where the metal may be vaporized and deposited over other parts of the tube. In the ignitron type of tube the arc is extinguished every cycle, and this problem is not of major importance.

The arc is established in the excitron by a jet of mercury propelled up from the cathode to contact an excitation anode, causing a temporary short circuit and arc formation. A magnetically controlled plunger operates the mercury jet when required, and the arc is maintained, once established, by the excitation anode circuit.

In the use of a control grid around the anode and suitable baffling the excitron is similar to ignitrons and other rectifiers.

BIBLIOGRAPHY

Hull, Dr. A. W., Gas-filled Thermionic Tubes, *J. A.I.E.E.*, Vol. 47, 798–803 (1928).
Hull, Dr. A. W., Hot-cathode Thyratrons, *Gen. Elec. Rev.*, Vol. 32, 213–223, 390–399 (April and July 1929).
Knowles, D. D., The Grid-glow Tube Relay, *Electric J.*, Vol. 25, 176–178 (April 1928).
Knowles, D. D., The Theory of the Grid-glow Tube, *Electric J.*, Vol. 27, 116–120 (February 1930).
Livingston, O. W., and H. T. Maser, Shield Grid Thyratrons, *Electronics*, Vol. 7, 114–116 (April 1934).
Morack, M. M., and H. C. Steiner, Sealed-tube Ignitron Rectifiers, *A.I.E.E. Trans.*, Vol. 61, 594–598 (1942).
Steiner, H. C., A. C. Gable, and H. T. Maser, Engineering Features of Gas-filled Tubes, *Elect. Eng.*, Vol. 51, 312–317 (May 1932).
Pike, O. W., Power Control through Grid-controlled Heavy Current Tubes, *Electronics*, Vol. 1, 85–87 (May 1930).
Slepian, J., and L. R. Ludwig, A New Method of Starting an Arc, *Elect. Eng.*, Vol. 52, 605–608 (September 1933).
Packard, D., and J. H. Hutchings, Sealed-off Ignitrons for Welding Control, *A.I.E.E. Trans.*, Vol. 56, 37–40, 66 (January 1937).
Germeshausen, K. J., and H. E. Edgerton, A Cold-cathode Arc-discharge Tube, *Elect. Eng.*, July 1936, pp. 790 ff.
Ingram, S. B., Cold-cathode Gas-filled Tubes as Circuit Elements, *A.I.E.E. Trans.*, Vol. 58, 342 ff (1939).
Arnott, E. G. F., Ignitor Characteristics, *J. Applied Phys.*, Vol. 12, 660–669 (September 1941).
Winograd, H., Development of Excitron-type Rectifier, *A.I.E.E. Tech. Paper*, 44–78 (March 1944).

X-RAY TUBES

By S. Reid Warren, Jr.

22. GENERAL PHYSICAL REQUIREMENTS

X-ray tubes may be divided into three general classes: (1) cold-cathode, gas-filled tubes; (2) hot-cathode, high-vacuum tubes; (3) cold-cathode, high-vacuum tubes. The last-named tube has been used to produce x-ray exposures of short duration (approximately 1 μsec) by field emission (see reference 8). Hot-cathode x-ray tubes are now used almost exclusively. Nevertheless a description of gas-filled x-ray tubes is of more than historical interest, since the phenomena of electric discharge in these tubes show clearly the limitations and the necessity for the careful design of all x-ray tubes.

The x-ray tubes employed by Roentgen and his successors in the field of x-ray research up to 1912 contained an anode, a cold cathode, and sometimes an electrode called the "anticathode." Only two electrodes are necessary, although it was alleged that the presence of the third electrode was accompanied by greater stability of the electric discharge through the tube. They are arranged in a glass bulb in the manner shown in Fig. 1. The tube is exhausted to a pressure of a few microns of mercury. The gas contains a number of free electrons. If there is a potential difference of the order of 40 to 80 kv across the electrodes of the x-ray tube, these free electrons are accelerated in the direction of the positive terminal or anode. They acquire sufficient velocity to remove one or more electrons from atoms of gas in their path. The ions formed by the ejection of these electrons from gas atoms have net positive charges. Under the action of the electric field these positive ions are accelerated toward the cathode. When the positive ions impinge upon the cathode, electrons are ejected from it. They in turn are attracted toward the anode and acquire a velocity depending upon the anode-cathode voltage. The electrons which impinge upon the anode at high velocity cause the generation of x-rays near the surface of the anode. As a result of the positive-ion stream passing toward the cathode and the electron stream passing to the anode, a current flows through

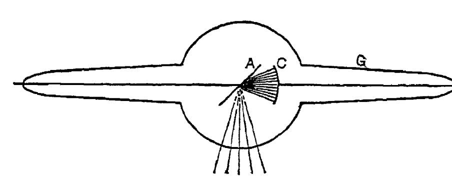


FIG. 1. Basic Elements of an X-ray Tube in Which Electrons are Ejected from a Cold Cathode C by Ionic Bombardment. G, the glass envelope. A, the anode disk of platinum, tungsten, or molybdenum. C, the cathode, made in the form of a segment of a sphere to focus the electrode stream upon a small area of the anode surface or focal spot.

the x-ray tube, which, in accordance with convention concerning the direction of current flow, is said to pass from anode to cathode.

The process involved in the generation of x-rays at voltages of the order of 100 kv is extremely inefficient. Of the total electrical energy supplied to an x-ray tube only about 0.01 per cent is transformed into useful x-rays. Of the remainder, the greatest part is dissipated as heat from the x-ray tube anode.

The current through an x-ray tube of this type cannot be controlled independently of the voltage. As the parts become heated during operation the gas pressure within the tube changes, often in an unpredictable manner, apparently influenced as much by the previous history of the tube as by the contemporary electrical input conditions. If the pressure within the tube decreases, a higher anode-cathode voltage is required to cause the flow of a given current. The x-rays generated at high voltage with low gas pressures are more penetrating than those generated at lower voltages with higher gas pressures. Low-pressure, cold-cathode tubes and the highly penetrating x-rays generated by operating them at high voltages are called *hard* tubes and *hard* x-rays. Tubes with relatively high pressure operated at low voltages are called *soft*, and the radiation so generated is called *soft* radiation. The terms *hard* and *soft* are purely qualitative.

In spite of ingenious attempts to make a practically useful cold-cathode x-ray tube, it has been used very little since the invention in 1913 of the hot-cathode x-ray tube by W. D. Coolidge and his co-workers.

Two important investigations during the period 1908–1913 contributed to the development of an x-ray tube in which the current can be varied independently of the voltage. Langmuir had investigated thoroughly the electronic emission from hot metallic filaments noted by Edison and earlier by Elster and Geitel. Briefly, to summarize this work, the following important findings are noted:

1. Unless the pressure of the gas within the tube is less than about 1 micron of mercury, positive ions produced in the same manner as those in the gas x-ray tube described above cause an important variation in the electron output of the filaments. At pressures lower than this the electrons emitted by the hot body are due apparently entirely to thermionic emission, and therefore their number depends only upon the cathode material and its temperature.

2. In general the introduction of various gases causes a considerable decrease in electron emission, particularly at low filament temperatures. This effect should not be confused with the effects due to ions formed in the added gas.

3. At very low gas pressures the thermionic current from a tungsten filament varies with the temperature in the following manner:

$$i = aT^2e^{-(b/T)}$$

T is the absolute temperature; i, the current in amperes per square meter of cathode surface; for tungsten $a = 6.02 \times 10^5$ amp per sq m per (degree)^2; $b = 52{,}400$ deg K. An expression similar to this was first developed by Richardson.

4. The operation of a tube with high vacuum is greatly influenced by a space charge consisting of a cloud of electrons surrounding the hot filament.

Investigators have found that x-rays are generated more efficiently by elements with high atomic number. Therefore they attempted to find a metal with the following characteristics:

1. High atomic number, to provide efficient x-ray generation.
2. High heat conductivity, to allow for dissipation of the heat generated at the x-ray tube anode.
3. Low vapor pressure, to assure stable operation.
4. High melting point, to provide sufficient capacity.
5. Ease of machining and relatively low cost.

X-ray tubes employed by workers who immediately followed Roentgen contained anodes of platinum. It was found subsequently that tungsten was better suited than platinum in all respects except that it was impossible, by means developed up to that time, to purify and to machine. A period of several years was required to devise methods for producing and working tungsten. The manufacture of uniform, standard x-ray tubes depends upon the carefully controlled processing of the component metal and glass parts. The method now generally used requires the removal and purification of an oxide of tungsten from the ore wolframite. The powdered oxide is formed into blocks and heated in the presence of hydrogen. After such treatment, the oxide is reduced to metallic tungsten; the bar contracts but remains extremely weak mechanically. It is, however, sufficiently strong so that bars about 2 cm square and 15 to 20 cm long may be supported at one end by a water-cooled copper electrode, while the other end rests in a pool of mercury which is also cooled. This system is surrounded by a jar through which hydrogen passes. An electric current is passed through the tungsten bar, heating it to about 2800 deg cent. Gradually the metal contracts and becomes mechanically stronger. It is then heated to red heat in air and swaged. This process consists of operating upon the tungsten bar with tools which pound it radially from several directions. By this treatment the tungsten becomes hard, homogeneous, and ductile. It is possible to continue the swaging process until the bar has been reduced to a diameter of a few millimeters. It may then be drawn into tungsten wire by means of diamond dies. The larger bars may be cut or polished by means of Carborundum. Solid tungsten anodes for x-ray tubes may be constructed from such bars. Smaller sections can be cut by means of thin (0.01 in.) rubber disks a few inches in diameter containing 240-grit Carborundum and operating, wet, at high speeds (9000 rpm).

W. D. Coolidge devised an x-ray tube like the one shown in Fig. 2. He found that it had the following characteristics:

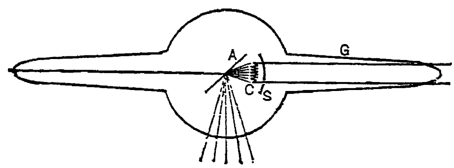


FIG. 2. Basic Elements of an X-ray Tube in Which Electrons are Ejected from a Hot Cathode *C* by Thermionic Action. *G*, the glass envelope. *A*, the anode disk, usually of tungsten. *C*, the cathode coil of tungsten wire. *S*, the metal focusing shield connected electrically to one of the cathode leads.

1. The tube allows current to pass only in the direction from anode to cathode.
2. The current through the x-ray tube is practically independent of the voltage applied to the anode and cathode for voltages of 30 to 200 kvp. The current may be varied by changing the temperature of the tungsten cathode.
3. The area upon the anode over which electrons impinge may be controlled by the cathode shield *S* and by the position of the cathode coil *C* within this shield.

X-ray tubes are employed for the following purposes:

Purpose	Voltage Limits	Current Limits
1. Generation of x-rays for the treatment of disease (x-ray therapy)	12–2000 kvp	0.1– 30 ma
2. The production of x-ray shadow pictures for medical diagnosis (roentgenography)	30– 110 kvp	10 –500 ma
3. Industrial roentgenography	30–2000 kvp	0.1–100 ma
4. X-ray diffraction	10– 100 kvp	2 –100 ma

Tubes for these various purposes differ considerably in structure. Some of the tubes now available are described in the following sections.

23. TUBES FOR X-RAY THERAPY

Roentgenologists require, for general x-ray therapy, tubes that will operate under the following conditions:

1. Tube voltages of 50 to 400 kvp and tube currents of 2 to 30 ma (average), supplied by constant-potential, full-wave pulsating, or half-wave pulsating high-voltage generators.
2. Exposures of 0.5 to 30 min duration are used.
3. Insulation to minimize the hazard of electrical shocks to patient and operator are essential.
4. X-rays must be confined to a cone the axis of which passes through the tube target and the part of the patient's body to be treated, to minimize the possibility of x-ray burn to the operator and patient.

The length of an x-ray tube depends upon the maximum anode-cathode voltage to which it may be subjected. The ends of the tube, to which the electrical connections are made, must be separated sufficiently to prevent sparkover. The electrical stress in the glass envelope must be sufficiently low to prevent puncture. The tube must be mechanically strong enough to support the anode and cathode assemblies rigidly. Therapy tubes, for use at 50 to 400 kvp, immersed in oil, are 0.2 to 1.5 m long. Formerly the glass envelopes were made of soft glass. Manufacturing methods devised recently have led to the use of a hard glass such as Pyrex. Pyrex has superior heat-resisting qualities, greater dielectric strength, and greater mechanical strength than soft glass. Air-cooled therapy tubes have a spherical glass bulb, 15 to 25 cm in diameter, surrounding the anode and cathode. This method of construction has the following advantages:

1. The sphere radiates more heat than a small cylindrical envelope.
2. The glass is far enough from the anode and cathode so that it is not stressed by strong electric fields caused by the applied anode-cathode voltage.

3. Stray electrons do not collect on the interior of the spherical surface in sufficient quantity to cause unstable operation or puncture of the tube.

In the so-called *Universal* tube, Fig. 3, a bar of molybdenum 0.5 cm in diameter is supported by glass at the end of the anode stem and extends into the tube, concentric with its axis, to a point about 5 cm from the tube center. At this end of the bar a solid tungsten block is fastened. This block is truncated at an angle of 45° with the tube axis. The minor axis of the elliptical face thus formed is 2 cm long. The minor axis of the target or focal spot is approximately 1.8 cm. Heat is dissipated chiefly by radiation. A spherical glass bulb is required to aid this heat dissipation.

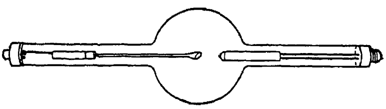


FIG. 3. *Universal* Therapy Tube, Solid Tungsten Anode. Air-cooled.

The tubes with spherical bulbs are difficult to mount and to shield by reason of their size. Modern therapy tubes are therefore designed for operation immersed in oil, with cooling of the anode by means of oil that is forced against the back of the anode by a pump. Special design of the electrodes, the use of hard-glass envelopes, and oil immersion result in tubes that are considerably smaller than tubes designed for mounting in air. An example of such a tube is shown in Figs. 4 and 5, designed to operate, in oil, at 250 kvp, 15 ma. Note the hollow cylindrical projection (hood) from the region of the tungsten target toward the cathode; this hood prevents secondary electrons from impinging on the inner surface of the glass envelope.

The cathode coil is usually a tungsten spiral which operates at 5 to 10 volts, 3 to 6 amp. The coil is supported and surrounded by a metal cup to shield the glass from secondary electrons and to focus the electron stream upon the target or focal spot of the anode. In a therapy tube a small focal spot or source of x-rays is not necessary. The focusing action of the cup is limited to preventing electrons from going past the anode face into the anode stem.

Tubes, such as that shown in Figs. 4 and 5, are often placed in a grounded metal tube head, supported by means of a tube stand that permits easy mechanical manipulation of the tube. High-voltage leads are flexible shockproof cables with a grounded external sheath. Figure 6 shows an apparatus of this kind being adjusted to make an x-ray film of a weld; therapy tube stands are essentially similar. The tube head and the rectangular cone extending from the head toward the weld are lead lined; thus the x-ray output is confined to a relatively sharply defined beam through the cone.

All x-ray tubes used for the routine treatment of deep-seated tumors are evacuated to a pressure of 10^{-2} micron of mercury or less. It is necessary not only to evacuate to this pressure while the anode and cathode are cold but also to heat the metal parts by means of high-frequency induction while the pumping proceeds. Before being sealed off, the tube is operated at an anode-cathode voltage and current slightly in excess of the tube rating to insure thorough degassing of the metal parts.

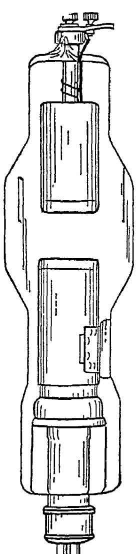


FIG. 4. X-ray Tube Designed for Operation, Immersed in Oil, at 250 kvp, 15 ma, Using Forced-oil Cooling. (Courtesy General Electric X-ray Corporation.)

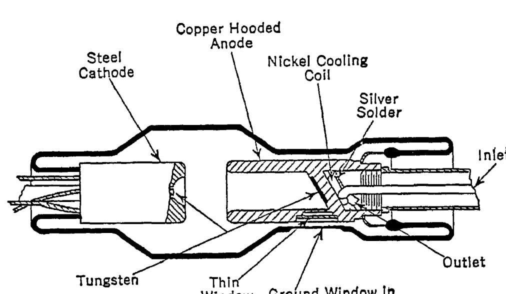


FIG. 5. Axial Section of Tube Shown in Fig. 4. (Courtesy General Electric X-ray Corporation.)

HIGH-VOLTAGE THERAPY TUBES. X-ray apparatus for therapy and for industrial radiography at tube voltages of 1000 and 2000 kvp are contained in steel tanks, in which air or "Freon" is maintained at a pressure equivalent to several atmospheres. The source of high-voltage, mounted in the tank with the x-ray tube, is a van de Graaff generator or a specially designed transformer operating at 180 cycles per second. The tubes are sealed, with the anode placed at one end of the cylinder; the cathode is placed at the other end. The anode is grounded, and the focal spot is cooled by water that flows over the end of the tube. A tube rated at 2000 kvp is shown in Fig. 7. This tube has a gold target; there is some evidence that the surface of the focal spot is molten during operation, although there is no evidence of the instability that might be expected to accompany this condition. Note that the tube is equipped with a series (180) of accelerating anodes. These are required to maintain uniform potential gradient along the tube and to minimize the effects of electrons' diverging from the beam. The magnetic focusing coil shown in Fig. 7 insures stable size, shape, and position of the focal spot on the gold target. The tube is approximately 9 ft long; the extension from the lowest accelerating anode to the target is 3 ft. The rating is 2000 kvp, 300 μamp.; the focal spot, with magnetic focusing, is approximately 2.5 mm in diameter.

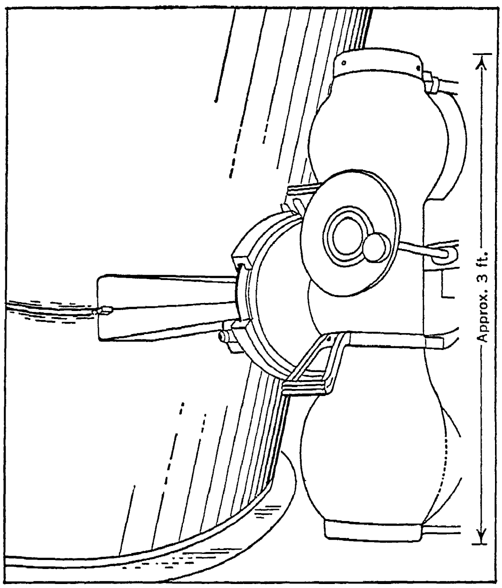


FIG. 6. 220-kvp Shockproof X-ray Tubehead and Tubestand. (Courtesy X-ray Division, Westinghouse Electric Corporation.)

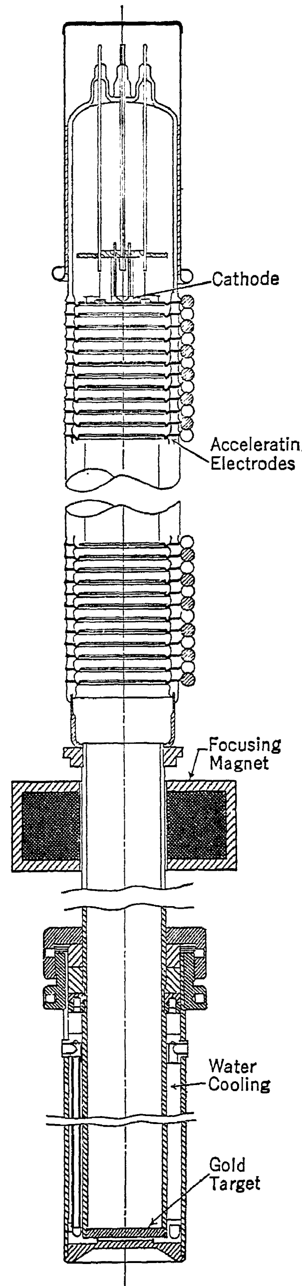


FIG. 7. Tube for Therapy and Industrial Radiography; Operating Voltage 2000 kvp. (Courtesy Machlett Laboratories.)

24. TUBES FOR MEDICAL ROENTGENOGRAPHY AND ROENTGENOSCOPY

To produce diagnostically useful roentgenograms of parts of the human body, x-ray tubes with the following characteristics are required:

1. The focal spot or source of x-rays must be as small as possible in order to obtain sharp shadow borders.
2. The technique of roentgenography prescribes x-ray tube voltages of 30 to 110 kvp, x-ray tube currents of 10 to 500 ma (average), and exposure times of $^1/_{30}$ to 30 sec.
3. The primary x-ray beam must be confined to a cone with the focal spot at the apex and the film at the base to protect the operator from x-ray burn.
4. Insulation of the tube and high-voltage leads is advisable but not necessary (except in dental roentgenography) since the focal spot-film distances generally used are greater than 60 cm.

Except for the requirement that the focal spot be small, the same principles are applied in designing roentgenographic tubes as those described above for therapy tubes. The need for minimizing the focal spot size having been prescribed, certain practical factors are found to oppose this decrease:

1. For any particular method of anode construction the energy dissipated at the focal spot during an exposure cannot exceed the value which will melt the tungsten target. The melting point of tungsten is 3370 deg cent.
2. Owing to the thickness of the part to be roentgenographed, the shadows of parts farthest from the film are more distorted than those near the film. The focal spot-film distance must be increased to decrease this effect. The energy supplied to the tube increases in proportion to the square of the focal spot-film distance. Therefore, the focal spot area must be increased in like proportion. In other words, if the focal spot-film distance is increased to reduce the magnification of shadows of parts at a distance from the film, the focal spot must be enlarged, the energy supplied to the tube must be increased, and the sharpness of the shadows remains unchanged.
3. Parts of the body are continually in motion. Roentgenograms of such parts must be made in as short a time as possible to minimize the unsharpness or blurring caused by motion during exposure. The minimization of exposure time requires an increase in focal spot size, all other factors remaining constant. For example, roentgenogram of the chest of a cadaver may be made with a tube having a focal spot less than 1 mm square. To make a roentgenogram of the chest of a normal adult requires a tube with a focal spot at least 3 mm square. Even this is a compromise, for no matter how short the exposure time some unsharpness is produced by motion.

Nearly all roentgenographic tubes now manufactured are constructed and used so that the line from the focal spot to the center of the film makes an angle of 10° to 45° with the anode face. The projection of the actual focal spot (the source of x-rays) upon a particular point in the film is called the effective focal spot area at that point. Figure 8 illustrates three types commonly used. The effective focal spot F_2 for the 45° tube is a circle 6 mm in diameter. The effective focal spots for the 20° and 10° tubes *having the same target areas as the 45° tube* are squares with sides of 3.7 mm and 2.6 mm, respectively. The maximum allowable exposure energies (for exposure times less than 0.2 sec) are equal for the three tubes, since the actual focal spot areas are equal. Therefore, for a given maximum rating the sharpness of the roentgenographic image increases as the angle between the focal spot and central x-ray beam decreases. Two limitations prescribe the minimum focal spot-film distances at which a tube may be used:

1. The intensity of x-rays emanating from the focal spot within a few degrees of the tangent to the anode face is less than the intensity of the central beam at the same distance from the tube. The intensity must be practically uniform over the area of the x-ray film, in the absence of absorbing objects between tube and film.
2. The effective focal spot size and the distortion vary over the film area. This variation must be minimized to insure that the roentgenogram is diagnostically useful over its entire area.

The design of roentgenographic tubes is determined in general by the criteria described above in article 23, "Tubes for X-ray Therapy." There are three requirements which differ qualitatively:

1. The anode-cathode voltages are less for roentgenography than for therapy.
2. It is extremely important to produce a small, uniform effective focal spot in roentgenographic tubes.
3. Exposure times for roentgenography are relatively short, requiring that the anode have a high heat capacity.

Modern roentgenographic tubes use the line or band focus shown in the 20° and 10° sketches, Fig. 8. The target is a disk of tungsten about 1.5 mm thick, usually with an area two or three times the actual focal spot area. The tungsten is embedded in a heavy copper bar. Suitable thermal contact between the tungsten and copper is difficult to obtain. The usual procedure involves fixing the tungsten disk at the bottom of a graphite crucible and pouring molten electrolytic copper over the disk. The casting is accomplished

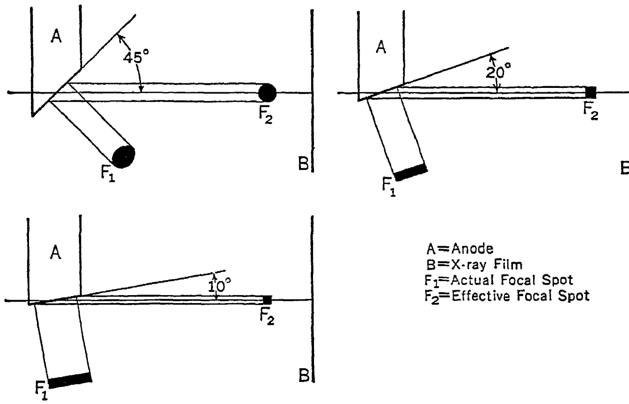


FIG. 8. Relations of Actual Focal Spots, F_1, and Effective Focal Spots, F_2, for Roentgenographic Tube Anodes. Actual focal spot areas of all three tubes are equal.

in an evacuated chamber to prevent the formation of any compounds that might prohibit the adherence of the copper to the tungsten.

With the type of construction just described, the maximum allowable exposure rating for exposures of 1 sec or less is 250 watts per square millimeter of actual focal spot area. For long exposures the rating depends upon the efficiency of the heat-radiating system, and therefore it is specified for each type of tube by the manufacturer. These data are given in graphical form, called tube rating charts.

Shock-proof roentgenographic tubes are constructed similarly to the therapy tube shown in Fig. 6.

To increase the rating for a focal spot of a given size, tubes have been developed with rotating anodes, the cathode stream being so directed that the focal spot is separated several millimeters from the tube axis. Different areas of the rotating anode therefore act as focal spot during the exposure. Two such tubes are shown in Figs. 9–12. The anode

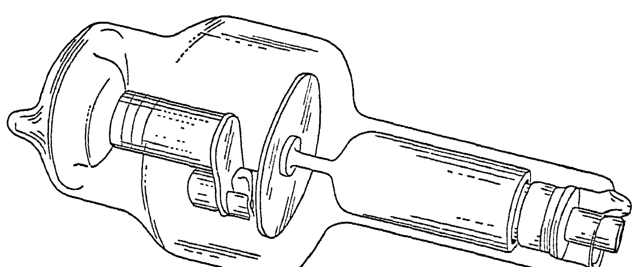


FIG. 9. Rotating-anode Roentgenographic Tube. Ratings (in oil-insulated tubehead): 2-mm focal spot, maximum voltage 100 kvp, 500 ma at 90 kvp for 1/30 sec, 400 ma at 80 kvp for 1/5 sec; 1-mm focal spot, maximum voltage 100 kvp, 200 ma at 90 kvp for 1/20 sec, 100 ma at 90 kvp for 3 sec. (Courtesy Machlett Laboratories.)

within the evacuated tube envelope operates as the rotor of a split-phase induction motor. These tubes are equipped with two cathodes; one produces an effective focal spot 1 mm square; the other, 2 mm square. The rotors operate at approximately 3000 rpm.

The tube shown in Figs. 9 and 10 is equipped with an anode comprising a tungsten disk approximately 6 cm in diameter and 2.3 mm thick. This disk is mechanically connected to the rotor of the driving motor by means of a short molybdenum shaft; the entire assembly is blackened to promote radiation of heat to the oil surrounding the tube when it is operating in its shockproof tube head. The balls of the bearing within the vacuum tube are coated with a thin, uniform film of silver, which reduces bearing friction. Samples of the manufacturer's ratings are listed below Fig. 9.

The anode structure of the tube shown in Figs. 11 and 12 consists of a tungsten cap backed by a heavy copper blackened cylinder, the purpose of which is to provide high heat storage and a large area for the radiation of heat to the oil surrounding the tube envelope. The ball bearings of this tube are lubricated by thin films of barium.

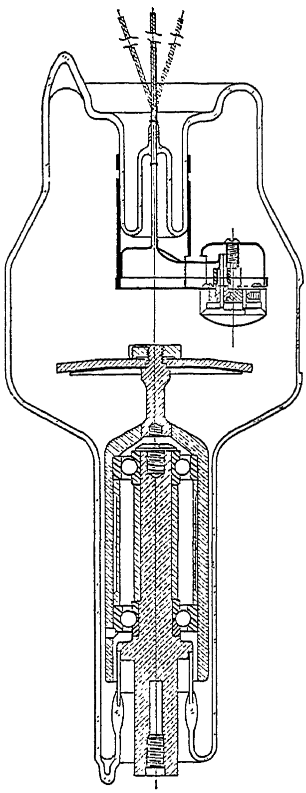


Fig. 10. Section of Rotating Anode Tube Shown in Fig. 9. (Courtesy Machlett Laboratories.)

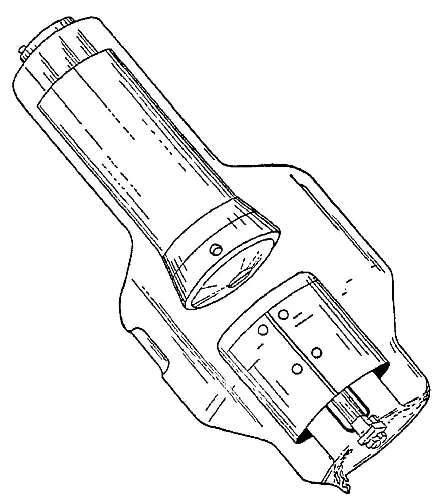


Fig. 11. Rotating-anode Roentgenographic Tube. Ratings (in oil-insulated tubehead): 2-mm focal spot, maximum voltage 100 kvp, 500 ma at 90 kvp for 1/30 sec, 400 ma at 80 kvp for 2/5 sec; 1-mm focal spot, maximum voltage 100 kvp, 200 ma at 94 kvp for 1/20 sec, 100 ma at 90 kvp for 5.6 sec. (Courtesy General Electric X-ray Corporation.)

Rotating-anode tubes are usually equipped with counters to record the number of exposures, and with control circuits that prevent high-voltage excitation of the tube unless the anode is rotating at rated speed (approximately 3000 rpm).

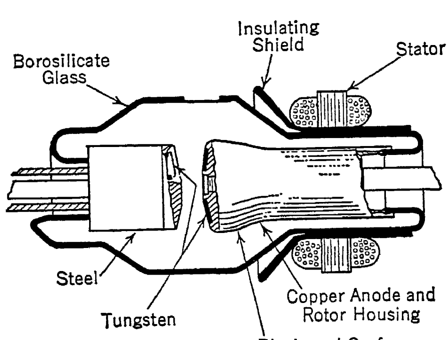


Fig. 12. Section of the Rotating-anode Tube Shown in Fig. 11. (Courtesy General Electric X-ray Corporation.)

Tubes used for roentgenoscopy (viewing of x-ray shadows by means of a fluorescent screen) are similar to the stationary-anode tubes described above, except that focal spots less than 2 mm square may be used, since the current seldom exceeds 10 ma. These tubes are often non-shockproof, since they are commonly mounted under a fluoroscopic table, out of reach of operator and patient.

Modern roentgenographic and fluoroscopic tubes should be used about 10–20 per cent below the manufacturers' ratings to guarantee a long useful life. Roentgenographic tubes, so operated, often can be used for 100,000 exposures or more.

25. TUBES FOR INDUSTRIAL ROENTGENOGRAPHY AND FLUOROSCOPY, AND FOR X-RAY DIFFRACTION

The use of x-rays (and of gamma rays) for the examination of commercial products has greatly increased, particularly during World War II; such examinations are quick and non-destructive. The following are representative of the diverse applications of this method of inspection: the examination for flaws in castings and welded joints; the inspection of packaged foods for foreign bodies; the final checking of manufacturing tolerances in such items as golf balls and multielectrode vacuum tubes; the inspection of packages for contraband; the inspection of citrus fruits suspected of being damaged by frost; the inspection of paintings and mummies.

X-ray tubes for industrial radiography and fluoroscopy are similar to those described in articles 23 and 24. The desire to use x-ray tube voltages up to 2 Mv for radiography of thick (as much as 0.3–0.4 m) steel was, in fact, a major stimulus to the development of the high-voltage tubes described in article 23 (Fig. 7). In all industrial radiographic tubes, regardless of voltage, as small a focal spot as possible must be produced to minimize the unsharpness of shadow borders, just as in medical roentgenography.

FIG. 13. Shockproof X-ray Diffraction Tube, with Grounded, Water-cooled Anode of Molybdenum, Copper, Cobalt, or Iron, and a Window of Beryllium. Ratings 50 kvp, 10–20 ma, continuous. (Courtesy Machlett Laboratories.)

Since industrial radiographic equipment may have to be moved about in a factory to use it near a heavy object to be inspected, design of equipment is based upon movability, flexibility of adjustment, and adequate protection of personnel from electric shock and excessive exposure to x-rays. Thus all industrial radiographic apparatus now manufactured is shockproof and shielded with lead or other x-ray-absorbing material, except in the direction of the useful beam.

The investigation of the structure of crystalline materials by means of x-ray diffraction has been developed from the early (1908–1915) theoretical analysis and experiments of von Laue, Friedrich, Knipping, W. H. and W. L. Bragg, and their colleagues. In recent years, x-ray diffraction apparatus suitable for routine analyses of crystal structure has been used in industry. X-ray tubes for this purpose consist of an anode of tungsten, molybdenum, copper, iron, or cobalt, a conventional incandescent tungsten cathode, and one or more windows of beryllium, which has an extremely low x-ray absorption coefficient.

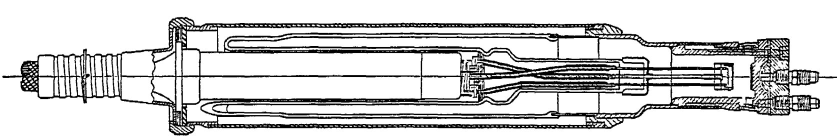


FIG. 14. Section of X-ray Diffraction Tube Shown in Fig. 13. (Courtesy Machlett Laboratories.)

The tubes operate at 10–50 kvp, 10–50 ma. A selective filter may be used to isolate the characteristic radiation from the anode, thereby producing a relatively intense, almost monochromatic beam of x-rays. A photograph of an x-ray diffraction tube and a cross-section of the tube are shown in Figs. 13 and 14.

BIBLIOGRAPHY

1. Kaye, G. W. C., *X-rays*, Longmans, Green (1926).
2. Coolidge, W. D., A Powerful Roentgen-Ray Tube with a Pure Electron Discharge, *Phys. Rev.*, Vol. 2, 409 (1913).
3. Langmuir, Irving, The Effect of Space Charge and Residual Gases on Thermionic Currents in High Vacuum, *Phys. Rev.*, Vol. 2, No. 6 (December 1913).
4. Bouwers, A., An X-ray Tube with Rotating Anode, *Physics*, Vol. 10, 125 (1930).
5. St. John, A., and H. R. Isenburger, *Industrial Radiography*, 2nd Ed. John Wiley (1943).
6. Siegbahn, M., *The Spectroscopy of X-rays.* Oxford University Press (1925).
7. Clark, G. L., *Applied X-rays*, 3d Ed. McGraw-Hill (1940).
8. Slack, C. M., and L. F. Ehrke, Field Emission X-ray Tube, *J. Applied Phys.*, Vol. 12, 165–168 (February 1941).
9. Machlett, R. R., and T. H. Rogers, A New Design for Increasing the Heat-dissipating Capacity of Rotating Anode Tubes, *Am. J. Roentgenology & Radium Therapy*, Vol. 48, 685–690 (November 1942).
10. Atlee, Z. J., Design and Application of X-ray Tubes, *Electronics*, Vol. 13, 26–30, 62, 64 (October 1940).

SECTION 5

ELECTRIC CIRCUITS, LINES, AND FIELDS

ELECTRIC CIRCUITS, LINES, AND FIELDS

THEORY OF LINEAR PASSIVE NETWORKS

By P. H. Richardson

The networks to be considered are assumed to consist of resistances, inductances, capacitances, and mutual inductances connected or coupled together in some manner. The problem is to determine the steady-state response of the network to an impressed voltage, or current, of any complexity. It is presupposed that the driving voltage has been impressed on the network at a time far enough in the past to have permitted any transients to die out.

To facilitate the solution of the problem the following assumptions are made:

1. The impressed voltage is periodic.
2. The network is *linear*. The values of the component elements are independent of the current through them.
3. The network is *passive*. There are no sources of energy interior to the network, and no energy is dissipated other than by the resistance elements of the network.

1. NON-SINUSOIDAL CURRENTS AND VOLTAGES

As a consequence of the linearity of the networks the coefficients of the differential equations are real constants. The equations express the equilibrium conditions which exist between the instantaneous driving voltages and the countervoltages in the circuit. For example, in the circuit of Fig. 1, consisting of R, L, and C $(= 1/D)$ in series, the differential equation is

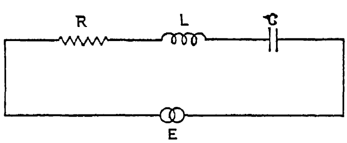


FIG. 1. Single-mesh Circuit

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + Dq = e \tag{1}$$

where q is the charge on the condenser, and $i = dq/dt$. If q_1 is the charge corresponding to an impressed voltage e_1, and q_2 is the charge corresponding to an impressed voltage e_2, it is evident that

$$L\frac{d^2q_1}{dt^2} + R\frac{dq_1}{dt} + Dq_1 = e_1$$

and

$$L\frac{d^2q_2}{dt^2} + R\frac{dq_2}{dt} + Dq_2 = e_2$$

By simple addition it follows that

$$L\frac{d^2(q_1 + q_2)}{dt^2} + R\frac{d(q_1 + q_2)}{dt} + D(q_1 + q_2) = e_1 + e_2 \tag{2}$$

Thus, for a linear network the *principle of superposition* holds. In other words, the current (or voltage) at any point flowing in response to several driving voltages acting together is the sum of the currents (or voltages) at that point which would flow in response to the driving voltages acting separately. This principle is of major importance in network analysis.

FOURIER'S THEOREM. A second important concept in the analysis of the steady state is Fourier's theorem, which states that any single-valued continuous periodic function can be expressed as an infinite series of sine waves. In particular, if $f(x)$ is a function which is finite in the interval from $-c$ to $+c$ and has only a finite number of discontinuities in that interval, then for any value of x in the interval

$$f(x) = \frac{a_0}{2} + a_1\cos\frac{\pi x}{c} + a_2\cos\frac{2\pi x}{c} + \cdots$$
$$+ b_1\sin\frac{\pi x}{c} + b_2\sin\frac{2\pi x}{c} + \cdots \tag{3}$$

where the coefficients $a_0, a_1, a_2, \cdots b_1, b_2, \cdots$ are determined as follows:

$$a_n = \frac{1}{c}\int_{-c}^{c} f(x)\cos\frac{n\pi x}{c}\,dx; \qquad b_n = \frac{1}{c}\int_{-c}^{c} f(x)\sin\frac{n\pi x}{c}\,dx$$

Or, equivalently,

$$f(x) = \frac{a_0}{2} + A_1 \sin\left(\frac{\pi x}{c} + \beta_1\right) + A_2 \sin\left(\frac{2\pi x}{c} + \beta_2\right) + \cdots \tag{4}$$

where

$$A_n = \sqrt{a_n^2 + b_n^2} \quad \text{and} \quad \beta_n = \tan^{-1}\left(\frac{a_n}{b_n}\right)$$

A proof of this theorem will be found in any standard book on calculus.

This theorem, taken in conjunction with the principle of superposition, makes it possible to obtain the response of a linear network to a periodic voltage of any complexity, provided that a solution is available for the case in which the driving voltage is a simple sinusoid. The complicated wave form of the impressed voltage is first resolved into its component sine waves, the solution of the equation is obtained for each component, and the solutions are added to obtain the final current or voltage required.

2. SINGLE-MESH CIRCUIT

Before proceeding to the general problem it will be profitable to examine in detail the response of the single-mesh circuit of Fig. 1. As already noted the differential equation is

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + Dq = E_0\epsilon^{j\omega t} \tag{5}$$

where the complex exponential form has been written for the impressed voltage. The notation is the usual one, where $\omega = 2\pi f$, f = frequency in cycles per second, E_0 is a constant, either real or complex, and $j = \sqrt{-1}$.

Assuming a solution of the form $q = q_0\epsilon^{pt}$ leads to the result that

$$(Lp^2 + Rp + D)q_0\epsilon^{pt} = E_0\epsilon^{j\omega t} \tag{6}$$

which is evidently a solution provided that $p = j\omega$. Then

$$q_0 = \frac{E_0}{Lp^2 + Rp + D}$$

and

$$i = I_0\epsilon^{pt} = \frac{E_0\epsilon^{pt}}{Lp + R + D/p} = \frac{e}{Z} \tag{7}$$

where $I_0 = E_0/Z$, $Z = R + Lp + D/p$, and $p = j\omega$.

Upon substituting $j\omega$ for p the result is the familiar one that

$$i = \frac{E_0}{Z}\epsilon^{j\omega t} \tag{8}$$

where $Z = R + jx$ and $x = (L\omega - D/\omega)$.

The use of the complex exponential form for the impressed voltage converts the differential equation at once to an algebraic equation. The resultant current is a complex exponential form of the same frequency as the impressed voltage. The impressed voltage is made up of two components in quadrature, and the resultant current is also made up of two currents in quadrature. Since the coefficients of eq. (5) are real, it is evident that the response to a real voltage is also real, while the current in response to an imaginary voltage is imaginary. Consequently, by the principle of superposition, the physical current flowing in response to a voltage $E_0 \cos \omega t$ is given by the real part of eq. (8). The imaginary part of the current, with the j discarded, is the physical current that flows in response to an impressed voltage $E_0 \sin \omega t$. Hence

$$i = \frac{|E_0|}{|Z|}\epsilon^{j(\omega t + \alpha - \beta)}$$

$$i_{\text{real}} = \frac{|E_0|}{|Z|}\cos(\omega t + \alpha - \beta) \tag{9}$$

and

$$i_{\text{imag}} = \frac{|E_0|}{|Z|}\sin(\omega t + \alpha - \beta)$$

where $Z = |Z| \underline{/\beta}$, $E_0 = |E_0| \underline{/\alpha}$, and $\beta = \tan^{-1}(x/R)$. The quantity $|E_0|$ is the maximum value of e, and similarly $|I_0|$ is the maximum value of i. By proper choice of units $|E_0|$ may be the rms voltage, and then $|I_0|$ is the rms current. Hence the ratio $|E_0/I_0|$ becomes the absolute value of the steady-state a-c impedance. The complex quantity $Z = e/i = E_0/I_0$ is defined as the complex steady-state impedance and may be studied as a function of $p = j\omega$.

3. THE COMPLEX FREQUENCY PLANE

The definition of the parameter $p = j2\pi f$ can be usefully extended for analytic purposes to situations in which f and p are complex. Suppose a voltage $E_0\epsilon^{pt}$ is assumed where $E_0 = |E_0| \underline{/\alpha}$ and $p = p_1 + jp_2$. Then

$$E_0\epsilon^{pt} = |E_0| \, \epsilon^{p_1 t}\epsilon^{j(p_2 t + \alpha)} \tag{10}$$

In accordance with the above discussion the physical voltage is taken as the real part of this, which is a sinusoidal oscillation with positive or negative damping depending on p_1. The corresponding steady-state physical current is obtained by dividing the complex voltage by the impedance and taking the real part of the result. This is a damped sinusoid of the same frequency and damping as the voltage.

Complex frequencies can be represented on a plane as shown in Fig. 2. The horizontal axis represents real values of p, and the vertical axis imaginary values of p, or real values

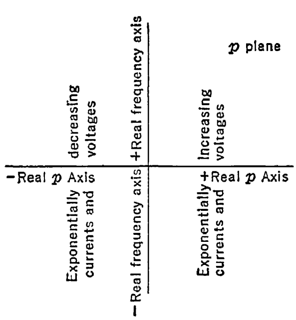


FIG. 2. Complex Frequency Plane

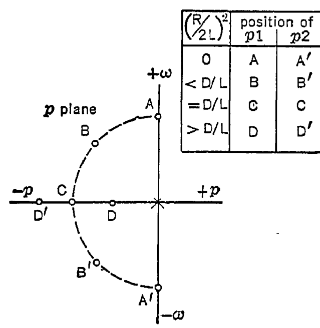


FIG. 3. Distribution of Zeros and Poles in Complex Plane

of frequency. In network analysis a distinction of primary importance is made between the right and left halves of the p plane, since on the left half-plane the voltages and currents correspond to functions which *decrease* exponentially with time, whereas on the right half-plane they *increase* exponentially with time. There is a close connection between the steady-state characteristics of a network and its transient response. Since a network whose transients increase with time is unstable, the characteristics of physical networks in the right half-plane are necessarily limited. There is no such distinction between the upper and lower halves of the plane.

As an illustration of this discussion consider the impedance of the single-mesh network of Fig. 1. Here the impedance

$$Z = \frac{L}{p}\left(p^2 + \frac{R}{L}p + \frac{D}{L}\right) \tag{11}$$

$$= \frac{L}{p}(p - p_1)(p - p_2)$$

where

$$p_1 = -\frac{R}{2L} + \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{D}{L}} \quad \text{and} \quad p_2 = -\frac{R}{2L} - \sqrt{\left(\frac{R}{2L}\right)^2 - \frac{D}{L}}$$

are the *zeros* of the impedance expression. They are seen to be real when $(R/2L)^2 \geq D/L$, while they are conjugate complex numbers when $(R/2L)^2 < D/L$. The locations of the zeros p_1 and p_2 are indicated on Fig. 3 as R/L is varied and D/L is held fixed. It should be noted that the values of p_1 and p_2 always have negative real parts; they are always in the left half-plane. The impedance becomes infinite, that is, it has a *pole*, when p is zero, as represented by the cross at the origin. A second pole occurs at infinity where the impedance again becomes infinite.

4. MESH EQUATIONS

The response of a complicated network of the type shown in Fig. 4 is determined by making use of the equilibrium conditions which must be satisfied by the instantaneous currents and voltages. There are several methods of writing the equations; one uses branch equations, a second mesh equations, and a third nodal equations. In writing branch equations the current in each branch, Z_{10}, Z_{12}, etc., is separately specified, and the sum of the instantaneous voltages in each branch is equated to the voltage applied to the ends of the branch. There are B such equations, where B is the number of branches. At each node, or junction point between branches, the sum of the currents entering the node must be equal to the sum of the currents leaving the node. Therefore, J relations of this kind can be found, where J is the number of junction points. If there are S separate parts (not conductively connected to one another, as mesh 7 in Fig. 4) only $J - S$ of these equations are useful, since, if the law of the conservation of charge is satisfied at all but one node in each of the S parts, it is automatically satisfied at the last one also. There are then $B + J - S$ independent equations.

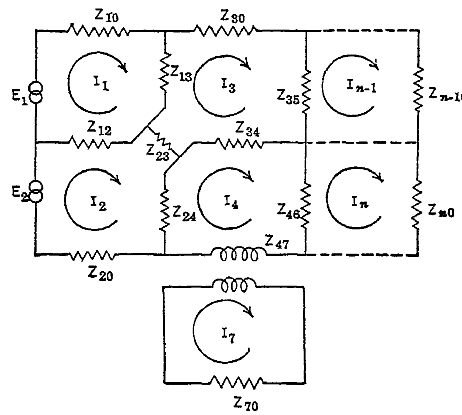


FIG. 4. Multi-mesh Network

The original branch equations include, in addition to the B branch currents, only differences in the node voltages. S of these can be arbitrarily assumed, and there will remain exactly $B + J - S$ unknowns to be determined.

By adding together the branch voltage equations around a complete loop, or mesh, and eliminating the superfluous branch currents by means of the nodal current equations, an equation is found for each mesh. A similar equation can be found for each of the N meshes of the network, where $N = B - J + S$.

A considerable simplification can be achieved by originally specifying only one current for each mesh of the network as shown in Fig. 4. By choosing circulating currents the nodal equations are automatically eliminated, as are the voltage differences which appear in the branch equations. The minimum number of mesh currents that can be used is, of course, $B - J + S$. For example, in Fig. 4 there are 15 branches, 10 junctions, and 2 parts, hence 7 meshes. It is important to note that a closed loop, consisting of a single branch with its two terminals coinciding, is considered to have one junction.

By adopting a set of conventions, regularity of notation is introduced into the equations. Those generally used are that all currents are assumed clockwise; branches common to two meshes a and b will carry as subscript the symbols of both meshes, as Z_{ab}; branches appearing in one mesh only will be designated Z_{a0}; and the sum of all the branches in a particular mesh will be designated as Z_{aa}. In computing Z_{aa}, the self-impedance of a mesh, all the self-inductance in the mesh is included but mutual inductances to other meshes are not.

Then for any linear network of N meshes a set of N equations is found as follows:

$$
\begin{aligned}
Z_{11}I_1 + Z_{12}I_2 + Z_{13}I_3 + \cdots Z_{1n}I_n &= E_1 \\
Z_{21}I_2 + Z_{22}I_2 + Z_{23}I_3 + \cdots Z_{2n}I_n &= E_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
Z_{n1}I_1 + Z_{n2}I_2 + Z_{n3}I_3 + \cdots Z_{nn}I_n &= E_n
\end{aligned}
\tag{12}
$$

where each of the Z's is of the form

$$Z_{ab} = pL_{ab} + R_{ab} + \frac{1}{p}D_{ab}$$

and p and $1/p$ represent differentiation and integration with respect to time.

The mesh equations developed thus far have represented the differential equations of the circuit. The set of differential equations is transformed to an identical set of algebraic equations by the assumption made in article 2, that each of the sinusoidal voltages and currents can be written as $E\epsilon^{j\omega t}$ or $I\epsilon^{j\omega t}$, where E and I are constants, $\omega = 2\pi f$, f = frequency in cycles per second, and $j = \sqrt{-1}$. This transformation results from the fact that differentiation and integration of $\epsilon^{j\omega t}$ replaces each p in eq. (12) by $j\omega$. The time factors $\epsilon^{j\omega t}$ appear on both sides of the equation and can be divided out. The Z's are then the complex self- and mutual impedances of the meshes. And the I's and E's are regarded as representing merely the constant coefficients in the more general expressions $I\epsilon^{j\omega t}$ and $E\epsilon^{j\omega t}$.

DRIVING POINT AND TRANSFER IMPEDANCES. The determination of any particular current flowing in response to a particular voltage is equivalent to the solution of a set of ordinary linear equations. The current in the first mesh flowing in response to the voltage $E_1\epsilon^{j\omega t}$, also in that mesh, is given by the method of determinants (see *Handbook of Engineering Fundamentals*, Eshbach, John Wiley) as

$$I_1\epsilon^{j\omega t} = \frac{\Delta_{11}}{\Delta} E_1\epsilon^{j\omega t} \tag{13}$$

where Δ is the determinant of the coefficients in the left-hand side of eq. (12) and Δ_{11} is the determinant obtained by omitting the first row and first column of Δ. The *driving point impedance* in the first mesh is by definition the ratio of the voltage to the current in eq. (13). Thus

$$Z = \frac{E_1}{I_1} = \frac{\Delta}{\Delta_{11}} \tag{14}$$

In similar fashion the current in the second mesh flowing in response to the voltage E_1 in the first mesh is given by

$$I_2\epsilon^{j\omega t} = -\frac{\Delta_{12}}{\Delta} E_1\epsilon^{j\omega t} \tag{15}$$

where Δ_{12} is the minor of Δ obtained by removing the first row and second column. The *transfer impedance* from the first to the second mesh is defined as

$$Z_T = \frac{E_1}{I_2} = -\frac{\Delta}{\Delta_{12}} \tag{16}$$

It should be noted also, since in a passive network $Z_{ab} = Z_{ba}$, that

$$Z_T' = \frac{E_2}{I_1} = -\frac{\Delta}{\Delta_{21}} = -\frac{\Delta}{\Delta_{12}} \tag{17}$$

5. NODAL EQUATIONS

An analogous system of equations can be set up in terms of driving currents at the nodes and nodal voltages. In this analysis the fundamental equations are conditions of current equilibrium. In the circuit of Fig. 5, I_1, I_2, I_3, and I_4 are the driving currents, and Y_{14}, Y_{12}, Y_{13}, etc., are the admittances (reciprocal impedances) of the various branches. Node 4 has been assumed at ground potential. At node 1, then,

$$Y_{14}E_1 + Y_{12}(E_1 - E_2) + Y_{13}(E_1 - E_3) = I_1$$

or (18)

$$Y_{11}E_1 - Y_{12}E_2 - Y_{13}E_3 = I_1$$

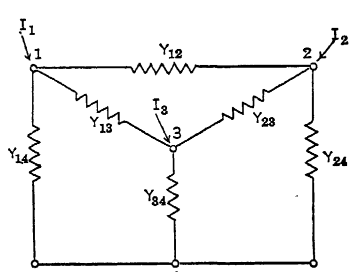


FIG. 5. Illustration for Method of Nodal Analysis

where $Y_{11} = Y_{14} + Y_{12} + Y_{13}$ is evidently the sum of the admittances of all the branches connected to 1 with all the other nodes connected together. Y_{11} is therefore a *self-admittance* analogous to the self-impedance in the mesh analysis. Similarly Y_{12} and Y_{13} are *mutual admittances* analogous to *mutual impedances*.*

* For a treatment of nodal analysis for circuits involving mutual inductance the reader is referred to Gardner and Barnes, *Transients in Linear Systems*, Vol. I.

In any conductively united network having J nodes a set of $J - 1$ independent equations of the above form can be written. The complete set becomes

$$\begin{aligned} Y_{11}E_1 - Y_{12}E_2 - \cdots - Y_{1n}E_n &= I_1 \\ -Y_{21}E_1 + Y_{22}E_2 - \cdots - Y_{2n}E_n &= I_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ -Y_{n1}E_1 - Y_{n2}E_2 - \cdots + Y_{nn}E_n &= I_n \end{aligned} \tag{19}$$

where the Y's are of the form

$$Y_{ab} = C_{ab}p + G_{ab} + \frac{1}{L_{ab}} \cdot \frac{1}{p}$$

and p and $1/p$ have the meanings previously ascribed to them.

A solution of the set of nodal equations to find the steady-state voltage corresponding to a given set of sinusoidal driving currents can be obtained by the processes already used. The *driving point admittance* Y between the first node and ground is defined as

$$Y = \frac{I_1}{E_1} = \frac{\Delta'}{\Delta_{11}'} \tag{20}$$

where the primes indicate that the Δ's refer to the set of equations (19).

Similarly the *transfer admittance* between the first and second nodes is defined as

$$Y_T = \frac{I_1}{E_2} = -\frac{\Delta'}{\Delta_{12}'} \tag{21}$$

In this case, also, since $Y_{ab} = Y_{ba}$,

$$Y_T' = \frac{I_2}{E_1} = -\frac{\Delta'}{\Delta_{21}'} = -\frac{\Delta'}{\Delta_{12}'} \tag{22}$$

One consequence of the analogy between the mesh and nodal equations is that the selection of one of the two methods of solution in any particular problem is entirely a matter of convenience. The symmetry of the two methods is further emphasized by the equivalence of Fig. 6, in which a constant-voltage generator in series with an impedance Z is shown as replaceable by a constant-current generator in parallel with an impedance Z.

A second consequence of the analogy leads to the *principle of duality* in network theory. The symmetry in the current and voltage methods of analysis includes the individual terms in the equations. The general term Z_{ab} of eq. (12) is replaced by Y_{ab} of eq. (19) if

$$L_{ab} = C_{ab},\ R_{ab} = G_{ab} \quad \text{and} \quad D_{ab} = \left(\frac{1}{C_{ab}}\right) = \frac{1}{L_{ab}}$$

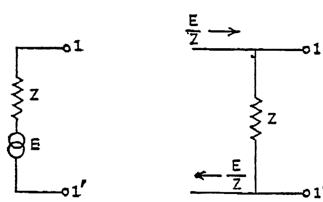


FIG. 6. Replacement of Constant-voltage Generator by an Equivalent Constant-current Generator

Consequently a set of nodal equations can be obtained identical with a given set of mesh equations by interchanging r and g, and L and C, wherever they appear. For every impedance function, therefore, there is a corresponding admittance function. If the mesh equations for one network correspond, term by term, with the nodal equations for another, the two networks are called *inverse* structures, or *duals*.

6. TWO-TERMINAL IMPEDANCES

The driving point impedance, or admittance, of a network can be expressed as the ratio of determinants whose elements are relatively simple functions of $p = j\omega$. In the mesh system the general impedance coefficient can be written as $Z_{ab} = L_{ab}p + R_{ab} + D_{ab}\,1/p$. Since the terms Δ, Δ_{11}, and Δ_{22} used in defining driving point impedances can be expressed as products of terms of this type, it follows that they are polynomials in p divided by some power of p. That is,

$$Z = \frac{A_m p^m + A_{m-1}p^{m-1} + \cdots + A_1 p + A_0}{B_n p^n + B_{n-1}p^{n-1} + \cdots + B_1 p + B_0} \tag{23}$$

Put in terms of the zeros and poles the expression is

$$Z = \frac{A_m(p - p_1)(p - p_2) \cdots (p - p_m)}{B_n(p - p_1')(p - p_2') \cdots (p - p_n')} \tag{24}$$

where ordinarily the p_a's are all different. Note that the p_a's and p_a''s are the roots of the polynomials in the numerator and the denominator. In special cases two or more zeros or poles may coincide. The zeros and poles may be thought of as corresponding to the resonances and antiresonances of purely reactive networks except that they may occur at complex frequencies.

RESTRICTIONS FOR PHYSICAL REALIZABILITY. A passive two-terminal or driving point impedance is subject to the following restrictions:

1. In terms of the frequency variable $p = j\omega$ the zeros and poles are either real or they occur in conjugate complex pairs.
2. The real and imaginary components are respectively even and odd functions of frequency on the real frequency axis.
3. None of the zeros and poles can be found in the right half of the p plane.
4. Zeros and poles on the real frequency axis must be simple.
5. The real component of the driving point impedance cannot be negative at real frequencies.

General methods for finding physical networks corresponding to any impedance function meeting these restrictions have been devised.* A method due to Brune depends on the fact that the minimum value of resistance, or conductance, at real frequencies is less than any value in the right half-plane. If the function is diminished by a real positive constant equal to the minimum value of the resistance, the remainder corresponds to a passive impedance having a zero resistance at a real frequency. This remainder is termed a *minimum resistance*, or *minimum conductance* expression. Similarly it can be shown that an impedance expression having zeros or poles on the real frequency axis can be diminished by the reactance, or susceptance, corresponding to its real frequency zeros and poles. An impedance having no poles at real frequencies is called a *minimum reactance* expression, while one having no zeros at real frequencies is called a *minimum susceptance* expression. If an impedance is both minimum resistance and minimum reactance, there is a unique relation between the resistance and reactance; if either is known at all frequencies the other can be determined.

NETWORKS OF PURE REACTANCES. If it is specified that the zeros and poles of an impedance expression occur at real frequencies, or imaginary values of p, the form of the impedance expression becomes

$$Z = \frac{kp(p^2 - p_2^2)(p^2 - p_4^2) \cdots (p^2 - p_m^2)}{(p^2 - p_1^2)(p^2 - p_3^2) \cdots (p^2 - p_{m-1}^2)} \tag{25}$$

where k is a positive real constant, while p_1^2, p_2^2, etc., are negative real quantities. Each of the factors represents a pair of zeros, or poles, at positive and negative real frequencies. The zeros and poles are restricted in that

$$-p_m^2 \geqq -p_{m-1}^2 \geqq - \cdots \geqq -p_2^2 \geqq -p_1^2 \geqq 0 \tag{26}$$

or the zeros and poles must alternate.

As written the impedance is specified as an inductive reactance at both zero and infinite values of frequency. To obtain complete generality it must be allowable to specify $p_1^2 = 0$ (which introduces a pole at $p = 0$), or that the factor $p^2 - p_m^2$ can be omitted (which leads to a zero at infinite frequency).

It has been demonstrated † that the impedance of eq. (25) corresponds to a physical network containing only inductances and capacitances. The network can be found either by representing the poles as antiresonant networks in series or by representing the zeros as resonant networks in parallel. For a detailed discussion of these impedances see Section 6, Article 17.

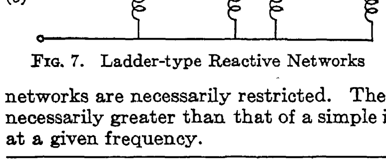


FIG. 7. Ladder-type Reactive Networks

The reactive networks can also be realized in other configurations of which Figs. 7*a* and 7*b* are typical. These are obtained as the result of continued fraction expansions in which the zeros and poles at zero and infinite frequency are removed alternately.

Because of the restrictions on the zeros and poles the impedance characteristics of reactive networks are necessarily restricted. The slope of the characteristic is always positive and necessarily greater than that of a simple inductance or capacity having the same reactance at a given frequency.

---

* Brune, *Journal of Mathematics and Physics, M.I.T.*, Vol. X, October 1931, pp. 191–235. Darlington, *Journal of Mathematics and Physics, M.I.T.*, Vol. XVIII, No. 4, September 1939, pp. 257–353.

† R. M. Foster, A Reactance Theorem, *B.S.T.J.*, April 1924, pp. 259–267.

NETWORKS OF RESISTANCES AND INDUCTANCES OR RESISTANCES AND CAPACITANCES. Very similar to the purely reactive networks which result when the zeros or poles are specified at real frequencies, a simple series of networks results when it is specified that the zeros and poles occur at imaginary frequencies, or real values of p. Again only two kinds of elements, either resistances and inductances or resistances and capacitances, are required to realize such impedances.

If the expression corresponds to a network of resistances and inductances, it is of the form

$$Z = \frac{kp(p - a_2)(p - a_4) \cdots (p - a_m)}{(p - a_1)(p - a_3) \cdots (p - a_{m-1})} \tag{27}$$

and $|Z|$ increases as p increases. If, on the other hand, the expression corresponds to a network of resistances and capacitances $|Z|$ decreases as p increases. An alternative form for Z is

$$Z = \frac{k(p - a_2)(p - a_4) \cdots (p - a_m)}{(p - a_1)(p - a_3) \cdots (p - a_{m-1})} \tag{28}$$

In both expressions the zeros and poles occur alternately and a_1, a_2, etc., are negative real quantities or zero.

As in the case of the reactive network the impedances may be represented in partial fraction form. For example, the impedance of eq. (28) can be expanded in the form

$$Z = \frac{D_1}{p - a_1} + \frac{D_3}{p - a_3} + \cdots + \frac{D_{m-1}}{p - a_{m-1}} + R \tag{29}$$

where $D_a = [(p - a_a)Z]_{p=a_a}$ and R is the resistance at infinite frequency. Each of the terms $D_a/(p - a_a)$ is identifiable with the parallel combination of resistance and capacitance, where $r_a = -D_a/a_a$ and $C_a = 1/D_a$.

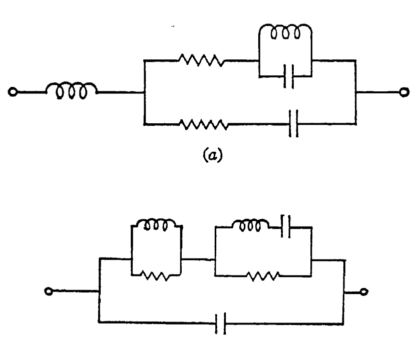
INVERSE OR RECIPROCAL IMPEDANCES. The duality between the impedance and admittance methods of analyzing a network suggests the possibility that to every network there corresponds an inverse. The requirement that the real part of an impedance be positive is merely another way of stating that the real part of the corresponding admittance be positive. Also, the restrictions on the zeros and poles are identical, so that the interchange of zeros and poles when an impedance is replaced by its reciprocal does not change the conditions for physical realizability. It follows, then, that, if a passive impedance is physically realizable, its reciprocal is also.

The reciprocal impedance for the structure of Fig. 8*a* is found, for example, as follows:

1. Each series connection is replaced by a parallel connection, and vice versa.
2. The individual resistances, inductances, and capacitances are respectively replaced by resistances, capacitances, and inductances in such a way that

$$R_1R_1' = \frac{L}{C'} = \frac{L'}{C} = R_0^2$$

The structural inverse of Fig. 8*a* is therefore given by 8*b*. This process is evidently not general since it considers only series and parallel connections. An extension of the method



(*a*) (*b*)

FIG. 8. Inverse Two-terminal Networks

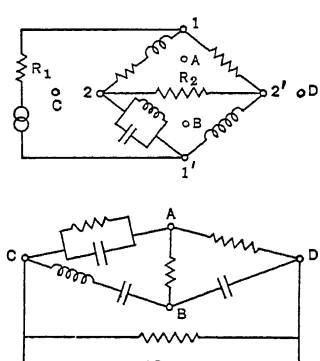


FIG. 9. Inverse Bridge Networks

depends on a consideration of the geometry of the network. The branches of the network are considered as lines between the junction points, dividing the plane of the diagram into areas. The process consists of interchanging areas and points. A new point is taken interior to each area and is connected to each similar point by a branch inverse to the branch separating the areas. The inverse of a bridge network is found by this process to be another bridge, as shown in Fig. 9. Even this process is not entirely general.*

COMPLEMENTARY IMPEDANCES. In addition to the inverse of a given impedance function one can also speak of its *complement*. The complement is defined by the requirement that the sum of the original impedance and its complement is a real constant. The complement of a passive impedance can be found if the prescribed impedance has no poles on the real frequency axis, and if the sum of the impedance and its complement is chosen at least as great as the maximum value of the resistance of the original impedance. The constant resistance combination of Fig. 10 represents a simple example of the relationship.

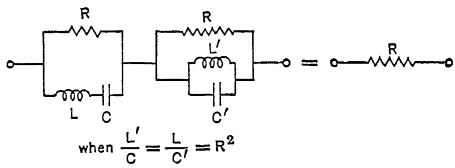


FIG. 10. Complementary Impedances in Series

7. FOUR-TERMINAL NETWORKS

The four-terminal network, or two-terminal pair, is a special form of general network of major importance. The external characteristics of the network are completely specified in terms of I_1, E_1, I_2, and E_2 of Fig. 11. A solution is obtained from eq. (12) with the assumption that all voltages except E_1 and E_2 are zero. Thus

$$I_1 = \frac{\Delta_{11}}{\Delta} E_1 - \frac{\Delta_{12}}{\Delta} E_2 \tag{30a}$$

$$I_2 = -\frac{\Delta_{12}}{\Delta} E_1 + \frac{\Delta_{22}}{\Delta} E_2 \tag{30b}$$

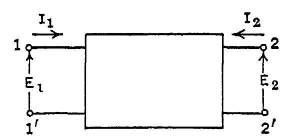


FIG. 11. General Four-terminal Network

Or, solving explicitly for E_1 and E_2, and noting that $\Delta\Delta_{1122} = \Delta_{11}\Delta_{22} - \Delta_{12}^2$

$$E_1 = \frac{\Delta_{22}}{\Delta_{1122}} I_1 + \frac{\Delta_{12}}{\Delta_{1122}} I_2 \tag{31a}$$

$$E_2 = \frac{\Delta_{12}}{\Delta_{1122}} I_1 + \frac{\Delta_{11}}{\Delta_{1122}} I_2 \tag{31b}$$

The currents of eqs. (30) are determined by the voltages E_1 and E_2 and the quantities Δ_{11}/Δ, Δ_{22}/Δ, and $-\Delta_{12}/\Delta$.

DRIVING POINT AND TRANSFER IMPEDANCES. The physical significance of the ratios is seen by successively setting E_2 and E_1 equal to zero. Thus

$\frac{\Delta_{11}}{\Delta} = Y_{s1}$ = admittance at terminals 1–1′ with 2–2′ shorted

$\frac{\Delta_{22}}{\Delta} = Y_{s2}$ = admittance at terminals 2–2′ with 1–1′ shorted

$-\frac{\Delta_{12}}{\Delta} = Y_{s12} = Y_{s21}$ = transfer admittance from either end with the opposite end shorted

Similarly, if I_2 and I_1 of eqs. (31a) and (31b) are successively set equal to zero,

$\frac{\Delta_{22}}{\Delta_{1122}} = Z_{01}$ = impedance at terminals 1–1′ with 2–2′ open

$\frac{\Delta_{11}}{\Delta_{1122}} = Z_{02}$ = impedance at terminals 2–2′ with 1–1′ open

$\frac{\Delta_{12}}{\Delta_{1122}} = Z_{012} = Z_{021}$ = transfer impedance from either end with the opposite end open

Thus the network may be described by either of these sets of three parameters, or by other sets of three parameters properly related to them.

* Foster, Geometrical Circuits of Electrical Networks, *Trans. A.I.E.E.*, June 1932.

The driving point impedances and admittances are subject to the same restrictions as any other two-terminal networks, if they are to correspond to physical networks. The transfer functions, however, differ in several important respects and require further consideration.

For the terminated network of Fig. 12 it can be shown that

$$E = \left(Z_1 + \frac{\Delta_{22}}{\Delta_{1122}}\right) I_1 + \frac{\Delta_{12}}{\Delta_{1122}} I_2 \qquad (32a)$$

$$E' = \frac{\Delta_{12}}{\Delta_{1122}} I_1 + \left(Z_2 + \frac{\Delta_{11}}{\Delta_{1122}}\right) I_2 \qquad (32b)$$

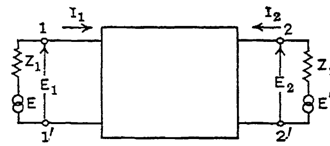


FIG. 12. Terminated Four-terminal Network

The response of the network evidently depends on the terminations Z_1 and Z_2, as well as on the network parameters themselves. Thus the driving point impedances become

$$Z_{1-1'} = \frac{\Delta + Z_2\Delta_{22}}{\Delta_{11} + Z_2\Delta_{1122}}; \qquad Z_{2-2'} = \frac{\Delta + Z_1\Delta_{11}}{\Delta_{22} + Z_1\Delta_{1122}} \qquad (33)$$

And the transfer impedances $Z_T = E'/I_1 = E/I_2$

$$Z_T = -\frac{1}{\Delta_{12}}[\Delta + Z_1\Delta_{11} + Z_2\Delta_{22} + Z_1Z_2\Delta_{1122}] \qquad (34)$$

The form of the transfer impedance expression is seen to be the ratio of determinants (since eq. [34] can be written $Z_T = -\Delta/\Delta_{12}$ if Δ is understood to be the network determinant including the terminations Z_1 and Z_2). The general impedance coefficient is $Z_{ab} = L_{ab}p + R_{ab} + D_{ab}(1/p)$ as before. In terms of the frequency variable $p = j\omega$, therefore, the transfer impedance is the ratio of polynomials in p and can be written in terms of its zeros and poles with a constant multiplier, just as in the case of driving point impedances. Thus

$$Z_T = \frac{A_m(p - a_1)(p - a_2) \cdots (p - a_m)}{B_n(p - b_1)(p - b_2) \cdots (p - b_n)} \qquad (35)$$

Since Z_T represents a transmission it is usually stated as a logarithm. And also, since the most efficient possible transmission between two impedances Z_1 and Z_2 is obtained if, first, the reactances of Z_1 and Z_2 are annulled, and then a transformer of optimum ratio is inserted between the two, this condition is used as a reference. Then the general transfer impedance is

$$Z_T = 2\sqrt{R_1R_2}\epsilon^\theta \qquad (36)$$

where

$$\epsilon^\theta = \epsilon^{A+jB} = \frac{k(p - a_1)(p - a_2) \cdots (p - a_m)}{(p - b_1)(p - b_2) \cdots (p - b_n)} \qquad (37)$$

and R_1 and R_2 are the real parts of Z_1 and Z_2. The a's and b's are the zeros and poles of Z_T, or the points of infinite gain or loss in terms of θ. It is evident from eq. (37) that two transfer impedances having the same zeros and poles can differ only by a constant loss or gain.

RESTRICTIONS FOR PHYSICAL REALIZABILITY. The restrictions which must be met if Z_T is to correspond to a physical network are as follows:

1. In terms of the frequency variable p both the zeros and the poles must be real or must occur in conjugate complex pairs.

2. The real and imaginary components are respectively even and odd functions of frequency.

3. The zeros must be located in the left half of the p plane; the poles may occur in any part of the plane.

4. The real part of $\theta = A + jB$ is positive at real frequencies; otherwise, the network serves as a source of power.

On comparing these restrictions to those given for two-terminal impedances two important differences are noted. First, the poles of the transfer impedance are not restricted to the left half-plane, and second, the real part of the transfer impedance may be negative. The previous restriction that the real part of a driving point impedance be positive is replaced by the new restriction that the real part of θ must be positive.

A structure which may be used to represent the general passive transfer function is shown in Fig. 13. The arms of the symmetrical lattice are assumed to be inverse such that $Z_x Z_y = R_2^2$. For this structure the ratio

$$\frac{E_1}{I_2} = Z_T = 2\sqrt{R_1 R_2}\epsilon^\theta \tag{38}$$

where

$$\epsilon^\theta = \frac{R_2 + Z_x}{R_2 - Z_x} \tag{39a}$$

or

$$Z_x = R_2 \frac{\epsilon^\theta - 1}{\epsilon^\theta + 1} \tag{39b}$$

It can be shown that Z_x represents a physical impedance as long as θ satisfies the restrictions listed above. In particular, Z_x will be physical as long as the transfer loss A is greater than zero at real frequencies. Consequently, in any case in which the minimum value of A is finite, the loss may be reduced by a constant corresponding to this minimum value. The reduced expression having zero loss at a real frequency is called a *minimum loss* or *minimum attenuation* function.

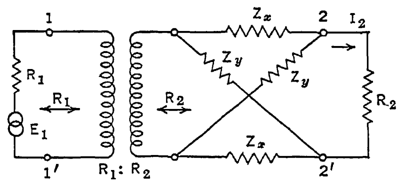


FIG. 13. Network Having Any Prescribed Passive Transfer Function

It is also possible to modify the transfer phase B without affecting the loss by introducing, or eliminating, terms corresponding to phase-shifting networks. Physical networks can be found (article 10) corresponding to the combinations of factors

$$\frac{p + a}{p - a} \quad \text{and} \quad \frac{(p + a + jb)(p + a - jb)}{(p - a - jb)(p - a + jb)} \tag{40}$$

Note that the poles in these expressions occur in the right half of the p plane, and that the zeros and poles are the negatives of one another. The absolute value of these expressions is evidently unity for all values of p, but the phase angle depends on p.

If the function Z_T contains a single pole on the positive real axis,

$$Z_T = \frac{F(p)}{p - a} = \frac{F(p)}{p + a}\left(\frac{p + a}{p - a}\right)$$
$$= Z_T'\left(\frac{p + a}{p - a}\right) \tag{41}$$

The transfer impedance Z_T is thus the product of a new transfer impedance Z_T' and a phase-shifting term, and can be shown to correspond to a new network of transfer impedance Z_T' in tandem with a simple phase network. A similar treatment is possible for a pair of complex poles in the right half-plane. The modified transfer impedance having no poles in the right half-plane is termed a *minimum phase* function. No further reduction can be made in the phase characteristic of such a function without at the same time affecting the loss characteristic.

NETWORK THEOREMS. Several useful theorems can be stated for the general linear network as follows:

The Compensation Theorem. If an impedance ΔZ is inserted in a branch of a network the resulting current increment produced at any point in the network is equal to the current that would be produced at that point by a compensating voltage $-I\Delta Z$ acting in series with the modified branch, where I is the current in the original branch.

The Reciprocity Theorem. If an electromotive force E of zero internal impedance applied between two terminals of a network produces a current I in some branch of the network, then the same voltage E acting in series with the second branch will produce the current I through the first pair of terminals shorted together. This follows from eqs. (30) since the short-circuit transfer admittance is the same from either end of the network.

Thévenin's Theorem. With respect to any pair of terminals considered as output terminals the network can be replaced by a branch having an impedance Z_{s2}, equal to the driving point impedance at these terminals, in series with an electromotive force E, equal to the open-circuit voltage across these terminals. An analogous theorem can be expressed in terms of the short-circuit current entering the output node of the network.

These results follow from eq. (30*b*) and the definitions of the open-circuit impedances and short-circuit admittances. Thus, if $E_2 = -I_2Z_2$,

$$I_2 = \frac{-E_1Y_{s12}}{Y_{s2}(Z_{s2}+Z_2)} = \frac{-E_1Z_{012}}{Z_{01}(Z_{s2}+Z_2)}$$

The open-circuit voltage at terminals $2 - 2'$ is $-I_2Z_2$ when $Z_2 \to \infty$, whence

$$I_2 = \frac{E_{20}}{Z_{s2}+Z_2} \tag{42}$$

The short-circuit current at terminals $2 - 2'$ is $-E_1Y_{s12}$, whence

$$E_2 = \frac{I_{2s}}{Y_2 + Y_{s2}} \tag{43}$$

EQUIVALENT QUADRIPOLES. A useful concept in network analysis is that of "equivalence." Two four-terminal networks, or quadripoles, are considered to be equivalent when the fundamental relations describing the behavior of the networks with respect to their input and output terminals are identical. Such equivalences can be expressed in terms of the network determinant and its minors, the open-circuit impedances, the short-circuit admittances, or any other set of three convenient and properly related parameters.

IMAGE PARAMETERS. An important and useful set of parameters is based on the idea that the terminations of the network be so related to the network itself that the impedances looking in both directions from the input terminals, or in both directions from the output terminals, are the same. Referring to Fig. 12 and eq. (33) this means that $Z_{1-1}' = Z_1$ and $Z_{2-2}' = Z_2$. The impedances $Z_1 = Z_{I_1}$ and $Z_2 = Z_{I_2}$ are called image impedances and are functions of the network itself, since

$$Z_{I_1} = \sqrt{\frac{\Delta\Delta_{22}}{\Delta_{1122}\Delta_{11}}} \quad \text{and} \quad Z_{I_2} = \sqrt{\frac{\Delta\Delta_{11}}{\Delta_{1122}\Delta_{22}}} \tag{44}$$

The third parameter necessary to characterize the network is called the image transfer constant and is defined as

$$\theta = \frac{1}{2}\log\frac{I_1^2Z_{I_1}}{I_2^2Z_{I_2}} \tag{45}$$

That is, θ is one-half the logarithm of the ratio of the volt-amperes entering the network to the volt-amperes leaving the network when it is terminated in its image impedances. In terms of the network determinant

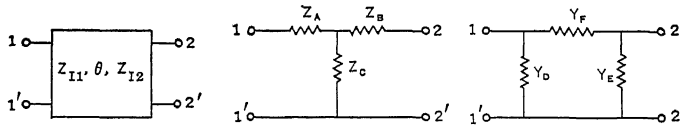
$$\theta = \frac{1}{2}\log\left[\frac{1+\sqrt{\Delta\Delta_{1122}/\Delta_{11}\Delta_{22}}}{1-\sqrt{\Delta\Delta_{1122}/\Delta_{11}\Delta_{22}}}\right] \tag{46}$$

The value of these parameters lies in the fact that they offer approximations to the actual network behavior and serve to relate the behavior of a network to that of a transmission line.

T AND π NETWORKS. The interrelation of several sets of network parameters is shown for two fundamental types of four-terminal networks in Fig. 14. The branches of the T and π networks equivalent to the general network are given in terms of the network determinant and its minors, the open- and short-circuit impedances, and the image parameters. These branch impedances and admittances themselves constitute complete sets of network parameters, even though they may not be physically realizable as two-terminal networks at all frequencies.

The T and π networks are important since they can be used to represent any quadripole for purposes of analysis and computation. They may evidently be used to represent any three-terminal network, or segment of a network, without loss of generality. The L-type network is considered to be a degenerate case of the T or π networks, and it is significant in that only two independent parameters are required to specify its behavior.

LATTICE OR BRIDGE NETWORKS. A network of major importance is the symmetrical balanced lattice shown in Fig. 15. It can be shown that this network is the most general of all symmetrical networks, and that any passive symmetrical four-terminal network, or quadripole, can be represented by a physically realizable lattice. Further, since the image impedance depends only on the product and the image transfer constant only on the ratio of the branch impedances, it is possible to control the transmission and impedance characteristics independently. See Section 6, article 23.



$$Z_A = \frac{\Delta_{22} - \Delta_{12}}{\Delta_{1122}} = Z_{01} - Z_{012} = \frac{Z_{I1}}{\tanh\theta} - \frac{\sqrt{Z_{I1}Z_{I2}}}{\sinh\theta} = \frac{Z_D Z_F}{Z_D + Z_E + Z_F}$$

$$Z_B = \frac{\Delta_{11} - \Delta_{12}}{\Delta_{1122}} = Z_{02} - Z_{012} = \frac{Z_{I2}}{\tanh\theta} - \frac{\sqrt{Z_{I1}Z_{I2}}}{\sinh\theta} = \frac{Z_E Z_F}{Z_D + Z_E + Z_F}$$

$$Z_C = \frac{\Delta_{12}}{\Delta_{1122}} = Z_{012} = \frac{\sqrt{Z_{I1}Z_{I2}}}{\sinh\theta} = \frac{Z_D Z_E}{Z_D + Z_E + Z_F}$$

$$Y_D = \frac{\Delta_{11} + \Delta_{12}}{\Delta} = Y_{s1} - Y_{s12} = \frac{Y_{I1}}{\tanh\theta} - \frac{\sqrt{Y_{I1}Y_{I2}}}{\sinh\theta} = \frac{Y_A Y_C}{Y_A + Y_B + Y_C}$$

$$Y_E = \frac{\Delta_{22} + \Delta_{12}}{\Delta} = Y_{s2} - Y_{s12} = \frac{Y_{I2}}{\tanh\theta} - \frac{\sqrt{Y_{I1}Y_{I2}}}{\sinh\theta} = \frac{Y_B Y_C}{Y_A + Y_B + Y_C}$$

$$Y_F = -\frac{\Delta_{12}}{\Delta} = Y_{s12} = \frac{\sqrt{Y_{I1}Y_{I2}}}{\sinh\theta} = \frac{Y_A Y_B}{Y_A + Y_B + Y_C}$$

$$Z_{I1} = \sqrt{\frac{\Delta\Delta_{22}}{\Delta_{1122}\Delta_{11}}} = \sqrt{Z_{s1}Z_{01}}$$

$$Z_{I2} = \sqrt{\frac{\Delta\Delta_{11}}{\Delta_{1122}\Delta_{22}}} = \sqrt{Z_{s2}Z_{02}}$$

$$\tanh\theta = \sqrt{\frac{\Delta\Delta_{11}}{\Delta_{11}\Delta_{22}}} = \sqrt{\frac{Z_{s1}}{Z_{01}}} = \sqrt{\frac{Z_{s2}}{Z_{02}}} = \sqrt{1 - \frac{Z_{s1}Z_{s2}}{Z_{s12}^2}} = \sqrt{1 - \frac{Z_{012}^2}{Z_{01}Z_{02}}}$$

$$Z_{s1} = Z_A + \frac{Z_B Z_C}{Z_B + Z_C} = \frac{1}{Y_D + Y_F}; \quad Z_{s2} = Z_B + \frac{Z_A Z_C}{Z_A + Z_C} = \frac{1}{Y_E + Y_F}$$

$$Z_{01} = Z_A + Z_C = \frac{1}{Y_D + \dfrac{Y_E Y_F}{Y_E + Y_F}}; \quad Z_{02} = Z_B + Z_C = \frac{1}{Y_E + \dfrac{Y_D Y_F}{Y_D + Y_F}}$$

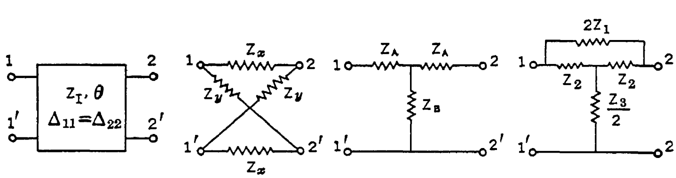
Symmetrical: ($\Delta_{11} = \Delta_{22}$, $Z_{01} = Z_{02}$, $Z_{s1} = Z_{s2}$)

$$Z_A = Z_B = Z_I \tanh\frac{\theta}{2} \qquad Y_D = Y_E = Y_I \tanh\frac{\theta}{2}$$

$$Z_C = \frac{Z_I}{\sinh\theta} \qquad Y_F = \frac{Y_I}{\sinh\theta}$$

$$Z_I = \sqrt{Z_A(Z_A + 2Z_C)} \qquad Y_I = \sqrt{Y_D(Y_D + 2Y_F)}$$

FIG. 14. T and π Networks Equivalent to a General Dissymmetrical Network



$$Z_x = \frac{\Delta_{11} - \Delta_{12}}{\Delta_{1122}} = Z_I \tanh\frac{\theta}{2} = Z_A = \frac{Z_1 Z_2}{Z_1 + Z_2}$$

$$Z_y = \frac{\Delta_{11} + \Delta_{12}}{\Delta_{1122}} = \frac{Z_I}{\tanh\theta/2} = Z_A + 2Z_B = Z_2 + Z_3$$

$$Z_I = \sqrt{Z_x Z_y} \qquad \tanh\frac{\theta}{2} = \sqrt{\frac{Z_x}{Z_y}}$$

FIG. 15. Lattice Network Equivalent to Any Symmetrical Network and to T and Bridged-T Networks

The lattice equivalent to the symmetrical T and bridged T structures of Fig. 15 is seen to be physical as long as the branch impedances of these structures are physical. The converse is not necessarily true, since each of these networks requires that the arms of the lattice contain a common impedance.

8. POWER TRANSFER

Transmission through a network is usually expressed as a logarithm with respect to a suitable reference condition.

Transition Loss. The condition for maximum power transfer from a generator to a load is indicated in Fig. 16. The series reactances $-X_1$ and $-X_2$ are inserted to annul the corresponding reactances of the generator and load impedances, and the ideal transformer of optimum ratio matches the resistance of the generator to that of the load. This structure is termed an *ideal transducer*, and the loss in power which is eliminated when it is inserted between a generator and a load is called the *transition loss* or the *transducer loss*. In decibels the

$$\text{Transition loss} = 10\log_{10}\frac{P_{20}}{P_2}$$

$$= 20\log_{10}|Z_1 + Z_2| - 10\log_{10}4R_1R_2 \quad (47)$$

where $P_{20} = E^2/4R_1$ is the reference power in the load (= available power) and $P_2 = E^2R_2/|Z_1 + Z_2|^2$ is the actual power in the load.

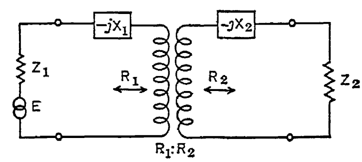


FIG. 16. Ideal Transducer

Insertion Loss and Phase. When a network is inserted between a sending impedance Z_1 and a receiving impedance Z_2 a change occurs in the current (or voltage) in the load. The ratio of the original load current I_{20} to the new load current I_2 is defined as the *insertion loss factor* or the *insertion factor*. In terms of the image parameters this becomes

$$\frac{I_{20}}{I_2} = \epsilon^{\Gamma} = \frac{k\epsilon^{\theta}}{k_1k_2S} \quad (48)$$

where

$$k_1 = \frac{2\sqrt{Z_1Z_{I_1}}}{Z_1 + Z_{I_1}} = \text{the sending end } \textit{reflection factor}$$

$$k_2 = \frac{2\sqrt{Z_2Z_{I_2}}}{Z_2 + Z_{I_2}} = \text{the receiving end } \textit{reflection factor}$$

$$k = \frac{2\sqrt{Z_1Z_2}}{Z_1 + Z_2} = \text{the } \textit{reflection factor} \text{ between } Z_1 \text{ and } Z_2$$

$$\theta = \text{the image transfer constant}$$

$$S = \frac{1}{\left[1 - \left(\frac{Z_1 - Z_{I_1}}{Z_1 + Z_{I_1}}\right)\left(\frac{Z_2 - Z_{I_2}}{Z_2 + Z_{I_2}}\right)\epsilon^{-2\theta}\right]} = \text{the } \textit{interaction factor}$$

The reflection factors represent modifications in the load current caused by reflections at the input and output junctions. The factor k, sometimes called the *symmetry factor*, represents a reflection factor that was eliminated when the network was inserted. Each of these factors involved only the ratio of two impedances, and each becomes unity when the impedances are equal.

The interaction factor, S, is a second-order effect which takes account of a wave reflected from the load back to the generator and then back to the load. It reduces to unity when either termination matches the image impedance adjacent to it, or when the attenuation of the network is high.

The insertion factor becomes equal to ϵ^{θ} when one of the terminating impedances and the image impedances are equal to one another.

The *insertion loss* (in decibels)

$$= 20\log_{10}\left|\frac{k\epsilon^{\theta}}{k_1k_2S}\right|$$

$$= 20\left[-\log_{10}\frac{1}{|k|} + \log_{10}\frac{1}{|k_1|} + \log_{10}\frac{1}{|k_2|} + \log_{10}\frac{1}{|S|} + \log_{10}|\epsilon^{\theta}|\right] \quad (49)$$

The first three terms, involving the reciprocals of the reflection factors, are called *reflection losses*. The fourth term is called the *interaction loss*, and the last term represents the real part of the image transfer constant in decibels.

The insertion phase shift is the phase angle of the current ratio given by eq. (48).

9. DISTORTION

When a voltage of complicated wave form is introduced into an electric circuit, a current will flow whose wave form will depend on that of the voltage and on the transmission characteristics of the circuit itself. It is frequently desired that the wave form of the current through a particular circuit element shall be the same as the wave form of the original voltage. If the complicated impressed voltage is regarded as a series of sine waves of various frequencies the conditions under which the output current will be a faithful reproduction of the input voltage may be stated as follows:

1. The response of the circuit must be the same for all frequency components present in the impressed voltage.
2. The relative phase relations of the various frequency components must not be altered.
3. The circuit must be linear.

FREQUENCY DISTORTION. When the first condition is not satisfied and the relative amplitudes of the various frequency components are altered, it is said that *frequency distortion* occurs. This type of distortion occurs when the voltage-current characteristic, that is the transfer impedance characteristic, is a function of frequency. Where the whole transmission apparatus (including any mechanical and acoustic portions) is considered, there are two methods by which the effect may be minimized. Each component part of the system may be so designed as to have its response independent of frequency, or some elements of the system may be designed to correct for the distortion introduced elsewhere. Both methods of design have been widely and successfully used, the choice for a particular case being decided usually by economic considerations.

DELAY DISTORTION. The requirement that the relative phases of the various components be unaltered is equivalent to the requirement that the time of transmission of the system be independent of frequency. To see this, consider a voltage consisting of a group of sine waves applied to the system. In complex notation

$$e = \sum_{k_1}^{k_2} E_k \epsilon^{j(\omega_k t + \theta_k)} \tag{50}$$

The received current will be

$$i = \sum_{k_1}^{k_2} \frac{E_k}{|Z_k|} \epsilon^{j(\omega_k t + \theta_k - \beta_k)} \tag{51}$$

where $Z_k = |Z_k| \underline{/\beta_k}$ is the transfer impedance of the circuit at each frequency $\omega_k/2\pi$.

It is evident that frequency distortion will be present if $|Z_k|$ is a function of frequency within the band of frequencies considered. If it be assumed that $|Z_k| = R$ and that $\beta_k = \omega t_1 \pm n\pi$, where R and t_1 are constants and $n = 0, 1, 2, \cdots$, then

$$i = \pm \frac{1}{R} \sum_{k_1}^{k_2} E_k \epsilon^{j[\omega_k(t - t_1) + \theta_k]} \tag{52}$$

The effect of the transfer impedance is to delay each component by an amount t_1 but to leave the relative phases unaltered. The change in sign which occurs when n is odd is usually not important.

If, on the other hand, $\beta_k = \omega_k t_1 + \sigma(\omega) \pm n\pi$, where $\sigma(\omega)$ is not a linear function of frequency, the wave form of the received current will be different from that of the impressed voltage even though the relative amplitudes of the current components are correct. (For a discussion of non-linear distortion see Non-linear Electric Circuits.)

10. CORRECTIVE NETWORKS

A corrective network, or equalizer, is a network inserted between a generator and a load such that the current in the load will vary with frequency in a predetermined manner. A loss, or attenuation, equalizer is one which is used to control the amplitude of the received current as a function of frequency without regard to phase relations. A phase equalizer is one which ideally introduces no loss but does introduce phase shift as a function of frequency.

In most cases it is sufficient to equalize only for changes in loss. If, however, both loss and phase equalization are required, it is usual first to equalize for loss, and then to correct the phase of the system plus the loss equalizers. The incidental loss characteristic introduced by the phase equalizer (because of power dissipated in the ideally reactive elements) is usually ignored. If necessary, this distortion is corrected by an additional loss equalizer designed as an integral part of the phase equalizer.

LOSS-PHASE RELATION. It is generally true that no unique relation can exist between the loss and phase characteristics of a four-terminal network. However, as noted in article 7, there is a unique relation between a given loss characteristic and the *minimum phase shift* that can be associated with it. This relation is of value in the design of corrective networks and feedback amplifiers, where it is necessary to control both loss and phase over wide frequency ranges.

For the minimum phase condition it is possible to derive a number of relations between loss and phase. One of the simplest is

$$\int_{-\infty}^{+\infty} \beta \, du = \frac{\pi}{2}(A_\infty - A_0) \tag{53}$$

where $u = \log \omega/\omega_0$, f_0 $(= \omega_0/2\pi)$ being an arbitrary reference frequency, β is the phase shift in radians, and A_0 and A_∞ are the losses in nepers at zero and infinite frequency respectively. This states that the area under the phase curve, when plotted on a logarithmic scale, depends only on the difference in the losses at zero and infinite frequency.

A second and possibly more useful relation is given by

$$\beta_0 = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{dA}{du} \log \coth \frac{|u|}{2} \, du \tag{54}$$

where β_0 represents the phase shift in radians at any arbitrary frequency f_0 $(= \omega_0/2\pi)$ and $u = \log \omega/\omega_0$. This result states that the phase shift at any frequency is proportional to the derivative of the loss, on a logarithmic frequency scale, at all frequencies. It involves an integration over the entire frequency spectrum. The function $\log \coth |u|/2$ is in the nature of a weighting function and is shown in Fig. 17. Its value is much larger near the point $\omega = \omega_0$ and tends to emphasize the effect of the loss characteristic in the immediate vicinity.

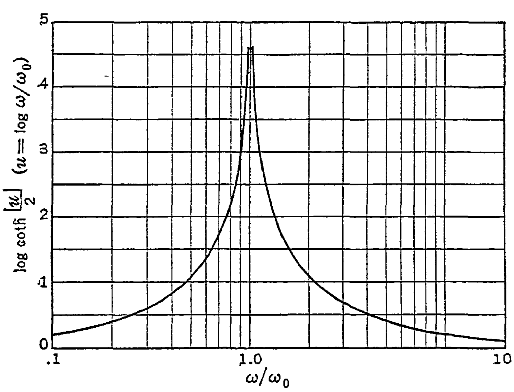


FIG. 17. Weighting Function in Loss-phase Formula

As an illustration of the utility of eq. (54) let it be supposed that $A = ku$, which describes a loss curve of constant slope on a logarithmic scale of $20k$ db per decade ($6k$ db per octave). The associated phase shift is readily found to be $k\pi/2$ radians. As a second example consider the discontinuous loss characteristic of Fig. 18. Here the loss is assumed to be zero below a specified frequency f_0 $(= \omega_0/2\pi)$, and has a constant slope of $6k$ db per octave above $\omega_0/2\pi$. The associated phase shift shown in Fig. 18 is symmetrical about the frequency f_0, at which point $\beta = k\pi/4$, and approaches the value $k\pi/2$ radians as frequency increases. At low frequencies the phase shift is substantially linear and is given by $\beta \doteq 2k\omega/(\pi\omega_0)$.

Since the phase characteristic corresponding to the sum of two loss characteristics is the sum of the two phase characteristics corresponding to the separate loss characteristics, it is possible to add a number of such simple characteristics together to simulate more complicated loss characteristics and to evaluate the corresponding phase shift. An example is furnished by Fig. 19, which shows a phase curve derived as the algebraic sum of three simple solutions of the type shown in Fig. 18. By proceeding in similar fashion it is possible to derive the phase shift corresponding to almost any loss characteristic without actually performing the integration indicated in eq. (54). In this connection it should be observed that, if both the loss and the corresponding phase can be specified at all frequencies, the problem of designing an equalizer having the inverse characteristics is immediately reduced to that of finding a two-terminal impedance for which both resistance and reactance are known. See article 7.

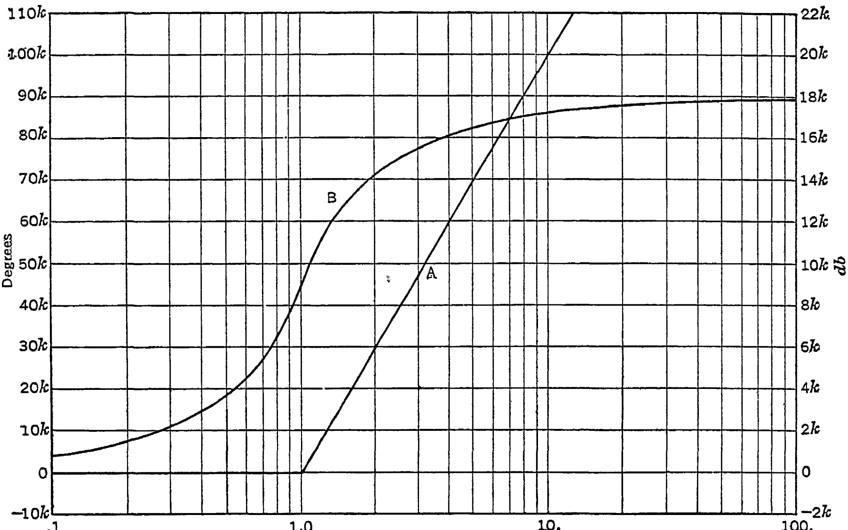


FIG. 18. Semi-infinite Slope of Attenuation (A) and Corresponding Phase Shift (B)

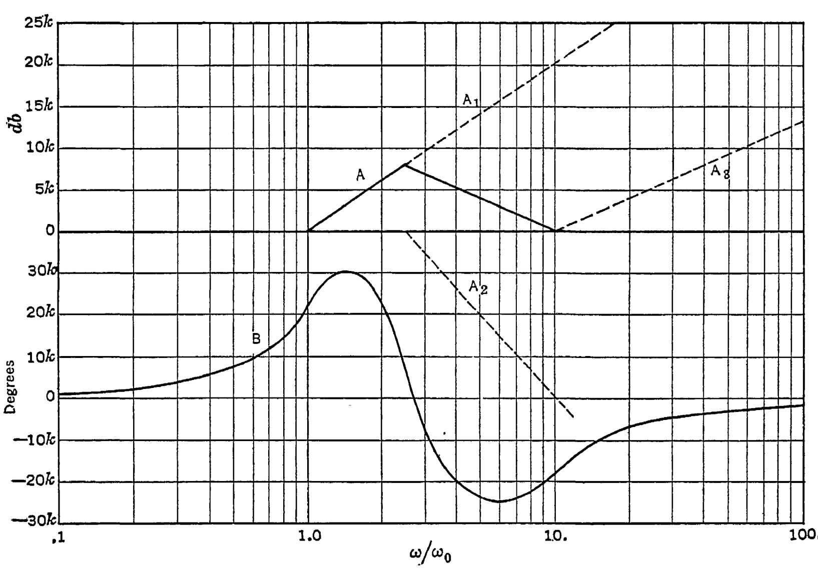


FIG. 19. Phase Curve Corresponding to Sum of Three Semi-infinite Attenuation Slopes

LOSS EQUALIZERS. The networks commonly used as loss, or attenuation, equalizers are shown in Fig. 20, which also gives the expressions for the insertion loss factor.

The networks shown are divided into classes based on their transmission and impedance properties. The simple series and shunt networks designated Ia and Ib are most useful for simple problems. Their transmission characteristics depend, of course, on both terminations.

The L-type networks designated as II*a* and II*b* have the same form for the insertion loss factor as I*a* and I*b*. They have the additional property, however, that the input impedance is equal to a constant, R_0, when they are terminated in R_0 on the output. Consequently, several sections can be operated in tandem without interaction, and the insertion loss of each section is independent of the generator impedance.

The symmetrical T, π, bridged T, and lattice networks shown as III*a*, III*b*, III*c*, and III*d* have insertion loss factors identical to the preceding network. However, they have constant-resistance image impedances. As a consequence the insertion loss factor has the form shown if either the generator or load impedance has the value R_0. Network III*c* is the most generally used because it requires fewer elements than any of the others.

The network shown as IV is a general constant-resistance lattice structure of which III*d* is a special case. This is the most general form of constant-resistance network, since any transmission characteristic which can be realized can be obtained with a structure of this form. See article 7. It is possible, therefore, to base the design of all equalizers on this structure, even though the network may be built in one of the other forms when such a conversion leads to a physical network.

The insertion loss factor, ϵ^θ, is somewhat more complicated for this network than for the others listed.

In terms of the admittance of the Z_1 arm

$$\epsilon^\theta = \frac{R_0Y_1 + 1}{R_0Y_1 - 1}$$

$$= \frac{R_0G_1 + 1 + jR_0B_1}{R_0G_1 - 1 + jR_0B_1} \quad (55a)$$

where $Y_1 = G_1 + jB_1$.

If G_1 is assumed to be constant with frequency, the network can be shown to behave as either a minimum phase or a non-minimum phase network as

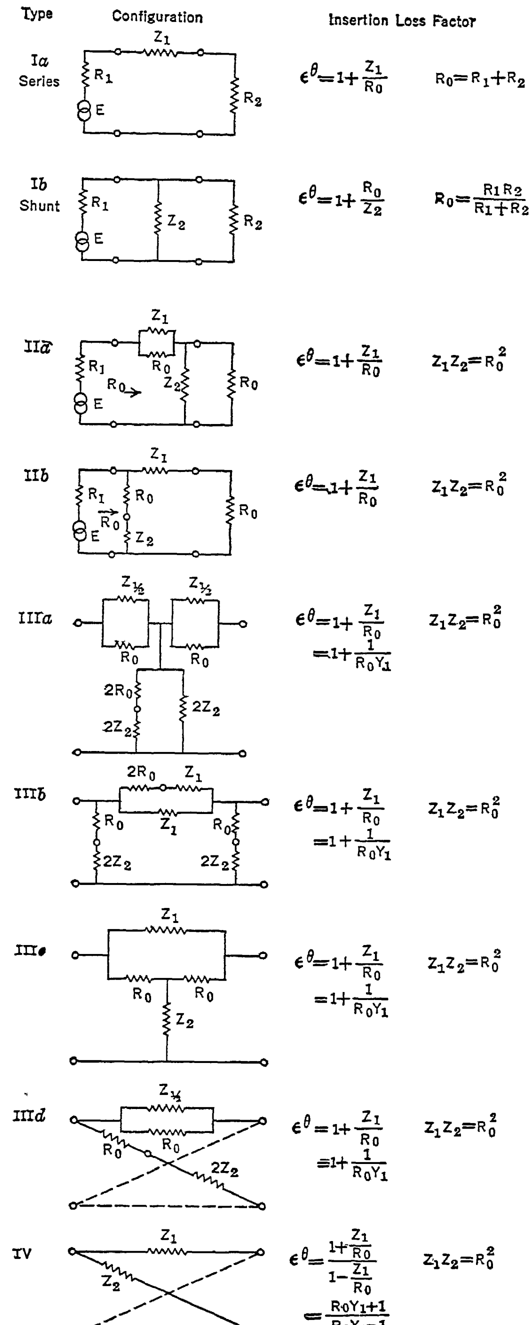


FIG. 20. Equalizer Configurations and Insertion Loss Factors

the product R_0G_1 is greater or less than unity. An explicit expression for Y_1R_0 in terms of θ is evidently

$$R_0Y_1 = \frac{\epsilon^\theta + 1}{\epsilon^\theta - 1} = \coth\frac{\theta}{2} \tag{55b}$$

A fairly general approach to equalizer design based on the constant-resistance lattice structure has been discussed by O. J. Zobel.* This method provides a systematic means of determining a network to satisfy a given loss requirement.

Referring to eq. (55a) it is a relatively simple matter to evaluate ϵ^θ corresponding to a given admittance Y_1. A more difficult problem is to determine Y_1 for a given set of values for α, since in a typical problem the corresponding value of β is usually unknown.

The design procedure is briefly as follows. The insertion power ratio is written as a function of frequency thus

$$\epsilon^{2\alpha} = \frac{a_0 + a_1\omega^2 + a_2\omega^4 + \cdots}{b_0 + b_1\omega^2 + b_2\omega^4 + \cdots} \tag{56}$$

where the a's and b's are real constants. A set of linear equations in the a's and b's is determined by assigning values to $\epsilon^{2\alpha}$ at specific frequencies. As many frequencies are selected as the number of coefficients to be determined. The solution of this set of equations gives the values of the a's and b's required, and the function $\epsilon^{2\alpha}$ is determined. The roots of the numerator and denominator of $\epsilon^{2\alpha}$ are then found in terms of $p^2 = -\omega^2$.

Since the function must lead to a physical network in order that it be useful the conditions must be satisfied that

1. $\epsilon^{2\alpha}$ be greater than 1.0 for all values of ω.
2. The roots of the numerator may not occur at real frequencies—otherwise $\epsilon^{2\alpha}$ would have a point of infinite gain.
3. The roots may be real or conjugate complex pairs.
4. The roots of the denominator may be anywhere in the p^2 plane, but those that fall on the negative real axis must be of even multiplicity—otherwise $\epsilon^{2\alpha}$ will jump from $+\infty$ to $-\infty$ in this region, and so violate restriction 1.

If these conditions are satisfied the design proceeds. If not, a change is required; either the assumed form for $\epsilon^{2\alpha}$ or the matching points must be altered.

To specify the network the function ϵ^θ is formed. The roots of $\epsilon^{2\alpha}$ are known in terms of p^2, and the corresponding roots in terms of p are readily available, since they occur in + and − pairs.

The roots of the numerator used to form ϵ^θ must be in the left-half of the p plane. The roots of the denominator are not necessarily so restricted. However, if this restriction is applied to them also, the solution for ϵ^θ will lead to a minimum phase network.

Utilizing the resulting function ϵ^θ the lattice arm impedance Z_1, or admittance Y_1, is determined from eq. (55b). The network which exhibits this impedance may be found in a variety of ways. A general method for solving this problem has been described by O. Brune.†

As an example of the design process consider that the function to be matched is

$$\epsilon^{2\alpha} = 1 + \omega^6$$

In terms of $p^2 = -\omega^2$ this becomes

$$\epsilon^{2\alpha} = 1 - p^6$$

The roots of $\epsilon^{2\alpha}$ are given by $p^2 = 1.0$ and $p^2 = -0.5 \pm j0.866$. In terms of p the roots having negative real parts are then $p = -1$ and $p = -0.5 \pm j0.866$, and

$$\begin{aligned}\epsilon^\theta &= (p + 0.5 - j0.866)(p + 0.5 + j0.866)(p + 1)\\ &= p^3 + 2p^2 + 2p + 1\end{aligned}$$

whence

$$R_0Y_1 = \frac{p^3 + 2p^2 + 2p + 2}{p^3 + 2p^2 + 2p}$$

Expanding this as a continued fraction by removing alternately poles and zeros the result is obtained that

$$R_0Y_1 = \frac{1}{p} + \frac{1}{2/p + p/(p + 1)}$$

The corresponding impedance of the series arm of the lattice is shown in Fig. 21 for unit impedance and unit frequency.

---

* O. J. Zobel, Distortion Correction in Electrical Circuits with Constant Resistance Networks, *B.S.T.J.*, Vol. VII, July 1928, pp. 438–534.

† *Journal of Mathematics and Physics, M.I.T.*, Vol. X, October 1931.

For many problems the analytic method of design which has been discussed is cumbersome and unsuitable. For example, if it be required to equalize a measured characteristic with only a fair degree of accuracy, the effort required to obtain a precise solution is not justified. In such cases a knowledge of the behavior of the simpler forms of two-terminal impedances can be usefully applied. The ability to visualize the frequency characteristic of a configuration of coils, condensers, and resistances is an essential part of the designer's equipment.

Probably the most useful configuration for design purposes, since it is one of the simplest, is that shown as IIIc in Fig. 20, in which the impedance Z_1 consists of a reactance shunted by a resistance. For this network

$$\epsilon^{\theta} = 1 + \frac{1}{R_0 Y_1}$$

$$= 1 + \frac{1}{R_0(G_1 + jB_1)} \tag{57}$$

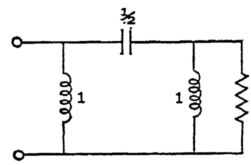


FIG. 21. Lattice Arm Impedance

where G_1 is a constant. Maximum loss occurs when $B_1 = 0$, while minimum loss ($\alpha = 0$) occurs when $B_1 \to \infty$. The design problem is reduced, once a selection has been made of the maximum loss required, to finding a suitable reactance, or susceptance, to match the loss curve over the required interval. From eq. (57) it is evident that

$$\epsilon^{2\alpha} = 1 + \frac{1 + 2R_0 G_1}{R_0^2(G_1^2 + B_1^2)} \tag{58}$$

whence

$$R_0 B_1 = \pm\sqrt{\frac{1 + 2R_0 G_1}{\epsilon^{2\alpha} - 1} - (R_0 G_1)^2} \tag{59}$$

For known values of $\epsilon^{2\alpha}$ corresponding values of B_1 can be found. A relatively slight effort is required to determine the required susceptance. The conditions for physical realizability are here merely that the zeros and poles occur at real frequencies, that they alternate, and that the reactance function behave as a simple coil or condenser at zero or infinite frequency.

In applying this process to build up complicated loss frequency characteristics as the sum of several equalizer sections the skill of the designer is evidenced by his ability to select easily realizable characteristics for the component sections.

PHASE EQUALIZERS. The constant-resistance lattice network IV of Fig. 20 bebecomes an "all-pass" network when the impedances Z_1 and Z_2 are specified as pure reactances. Then

$$\epsilon^{\theta} = \epsilon^{\alpha + j\beta} = \frac{R_0 + jX_1}{R_0 - jX_1}$$

$$\epsilon^{2\alpha} = 1.0 \quad \text{and} \quad \tan\frac{\beta}{2} = \pm\frac{X_1}{R}$$

There are two basic networks of this type, which are shown in Fig. 22. For the first of these (Type I)

$$\epsilon^{\theta} = \frac{R_0 + Lp}{R_0 - Lp} = -\frac{p + R_0/L}{p - R_0/L} \tag{60}$$

The zero of this expression is real and negative, and the pole is real and positive. In the p plane they occur symmetrically on either side of the origin. The phase curve corresponding to this network is shown in Fig. 22; the critical frequency $\omega_0 = R_0/L$.

In the more complicated network shown as Type II in Fig. 22

$$\epsilon^{\theta} = \frac{R_0 + \dfrac{Lp}{1 + LCp^2}}{R_0 - \dfrac{Lp}{1 + LCp^2}} = \frac{(p - p_1)(p - p_3)}{(p - p_2)(p - p_4)} \tag{61}$$

where

$$p_1 = -p_4 = \frac{\omega_0}{2}\sqrt{\frac{L}{C}}\frac{1}{R_0}\left[-1 + \sqrt{1 - \frac{4R_0^2 C}{L}}\right]$$

$$p_3 = -p_2 = \frac{\omega_0}{2}\sqrt{\frac{L}{C}}\frac{1}{R_0}\left[-1 - \sqrt{1 - \frac{4R_0^2 C}{L}}\right]$$

and

$$\omega_0 = \frac{1}{\sqrt{LC}}$$

When $4R_0^2C/L \leqq 1.0$ the zeros and poles are real. The zeros occur in the left half of the p plane, and the poles are in the right half-plane. The network is, therefore, equivalent to two of the simple types in tandem.

When $4R_0^2C/L > 1.0$ the zeros and poles are complex. Again the zeros are in the left half-plane and are conjugate complex numbers. The poles are again in the right half-plane and are symmetrically disposed about the origin. As would be expected a wide variety of phase characteristics can be obtained. Some typical curves are shown in Fig. 22 plotted against ω/ω_0. Note that, when $4R_0^2C/L = 1.0$, the phase shift is exactly twice that of Type I.

Increasing the complexity of the reactances in the lattice arms merely introduces additional zeros and poles in the expression for ϵ^θ. If the network is to continue to be an all-pass network, the zeros and poles must occur in pairs having characteristics similar to

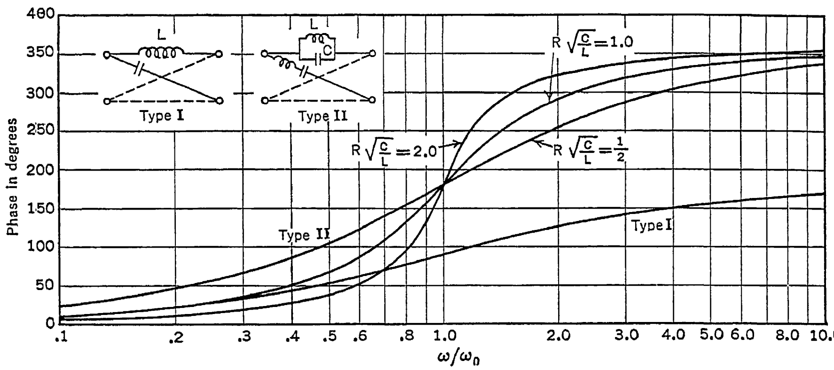


FIG. 22. Typical Phase Characteristics of All-pass Networks

those noted above. Consequently, it is possible to break down a more extensive set of zeros and poles into groups, each group corresponding to a physical network of either Type I or Type II. A number of these simpler networks in tandem will provide characteristics exactly similar to those of the more complicated lattice.

BIBLIOGRAPHY

Bode, H. W., *Network Analysis and Feedback Amplifier Design.* D. Van Nostrand (1945).
Gardner and Barnes, *Transients in Linear Systems*, Vol. I.
Brune, O., *Journal of Mathematics and Physics, M.I.T.*, Vol. X, October 1931, pp. 191–235.
Darlington, S., *Journal of Mathematics and Physics, M.I.T.*, Vol. XVIII, No. 4, September 1939, pp. 257–353.
Foster, R. M., A Reactance Theorem, *B.S.T.J.*, April 1924, pp. 259–267.
Foster, R. M., Geometrical Circuits of Electrical Networks, *Trans. A.I.E.E.*, June 1932.
Guillemin, E. A., *Communication Networks*, Vol. II.
Zobel, O. J., Distortion Correction in Electrical Circuits, *B.S.T.J.*, Vol. VII, July 1928, pp. 438–534.

RECURRENT NETWORKS

By P. H. Richardson

Early work on transmission networks was from the viewpoint of wave propagation in uniform media. Later work introduced the methods of particle dynamics, and networks were treated as vibrating systems. In treating the problem of *recurrent networks*, that is, the problem of a number of similar networks in tandem, the terminology of the earlier method is most useful. As might be expected, the problem is readily handled in terms of equivalent line parameters such as the image parameters previously defined.*

* Solutions of the general problem in terms of both image and iterative parameters are given by E. A. Guillemin, *Communication Networks*, Vol. II, pp. 163–175. In the case of symmetrical networks the two sets of parameters are identical.

11. SYMMETRICAL NETWORKS

CURRENT AND VOLTAGE RELATIONS. Consider the tandem combination of two symmetrical networks shown in Fig. 1. The image impedance is assumed to be the same for both structures, while the transfer constants θ_1 and θ_2 are different. To find the voltage

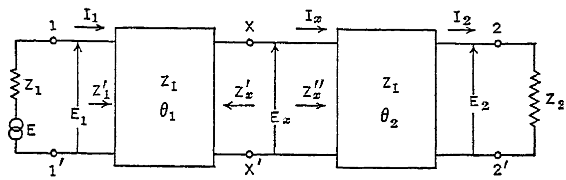


FIG. 1. Tandem Combination of Symmetrical Networks

at the junction x–x' replace the generator and network to the left of the junction by an equivalent Thévenin generator. Then the open circuit voltage at x–x'

$$E_{ocx} = \frac{EZ_I}{\left(Z_1 + \dfrac{Z_I}{\tanh\theta}\right)\sinh\theta} \tag{1}$$

The voltages E_x and E_1 are evidently given by

$$E_x = \frac{E_{ocx}}{1 + (Z_x'/Z_x'')} \quad \text{and} \quad E_1 = \frac{E}{1 + (Z_1/Z_1')} \tag{2}$$

The impedances Z_1', Z_x', and Z_x'' are given by the relations

$$Z_1' = Z_I\left[\frac{1 + \rho_2\epsilon^{-2(\theta_1+\theta_2)}}{1 - \rho_2\epsilon^{-2(\theta_1+\theta_2)}}\right] \tag{3a}$$

$$Z_x' = Z_I\left[\frac{1 + \rho_1\epsilon^{-2\theta_1}}{1 - \rho_1\epsilon^{-2\theta_1}}\right] \tag{3b}$$

$$Z_x'' = Z_I\left[\frac{1 + \rho_2\epsilon^{-2\theta_2}}{1 - \rho_2\epsilon^{-2\theta_2}}\right] \tag{3c}$$

where $\rho_1 = \dfrac{Z_1 - Z_I}{Z_1 + Z_I}$ and $\rho_2 = \dfrac{Z_2 - Z_I}{Z_2 + Z_I}$ are the *reflection coefficients* at junctions 1–1′ and 2–2′ respectively.

Substituting in eq. (2) it can be demonstrated that

$$E_x = E_1(A_1\epsilon^{-\theta_1} + B_1\epsilon^{\theta_1}) \tag{4}$$

$$I_x = \frac{E_x}{Z_x''} = \frac{E_1}{Z_I}(A_1\epsilon^{-\theta_1} - B_1\epsilon^{+\theta_1}) \tag{5}$$

where

$$A_1 = \frac{1}{1 + \rho_2\epsilon^{-2(\theta_1+\theta_2)}} \quad \text{and} \quad B_1 = \frac{\rho_2\epsilon^{-2(\theta_1+\theta_2)}}{1 + \rho_2\epsilon^{-2(\theta_1+\theta_2)}} \tag{6}$$

Note that A_1 and B_1 depend only on the sum of the two transfer constants and the reflection coefficient at the output junction.

INCIDENT AND REFLECTED WAVES. The term $E_1A_1\epsilon^{-\theta_1}$ is called the *incident component* of the voltage, and $E_1B_1\epsilon^{\theta_1}$ is called the *reflected component* of the voltage. Likewise the terms $(E_1/Z_I)A_1\epsilon^{-\theta_1}$ and $-(E_1/Z_I)B_1\epsilon^{\theta_1}$ are called the incident and reflected components respectively of the current at the junction x–x'. Note that if $Z_2 = Z_I$, that is, if the impedances are *matched* at the output, there is no reflected component since $B_1 = 0$. Also, if $\theta_1 + \theta_2$ has a large positive real part, B_1 becomes very small, and again the reflected component vanishes. The voltage ratio and the current ratio are both equal to $\epsilon^{-\theta_1}$, when the reflected "wave" is zero.

IMPEDANCE RELATIONS. The pair of networks having like image impedances behave as regards the input and output meshes as though they together constituted one network of image impedance Z_I and transfer constant $\theta_1 + \theta_2$. When the output impedance matches Z_I, $\rho_2 = 0$ and the impedances Z_1' and Z_x'' are both equal to Z_I. Note also, if the real part of $\theta_1 + \theta_2$ is large, the input impedance again is equal to Z_I.

A relation which is frequently useful in design problems can be obtained from eq. (3). The reflection coefficient at the input terminals

$$\rho_1' = \frac{Z_1' - Z_I}{Z_1' + Z_I} = \rho_2 \epsilon^{-2(\theta_1+\theta_2)} \tag{7}$$

Thus the reflection coefficient at the input terminals, ρ_1', is simply related to the reflection coefficient at the output terminals. If $\theta_1 + \theta_2$ is a pure imaginary, then $|\rho_1'| = |\rho_2|$. Also, if α is the real part of $\theta_1 + \theta_2$, then $|\rho_1'| = |\rho_2|\epsilon^{-2\alpha}$.

12. UNIFORM LINES—NETWORKS WITH DISTRIBUTED CONSTANTS

When the physical dimensions of a network are comparable to the wavelength of the electric current flowing in it, account must be taken of the fact that the series resistance and inductance of each wire, and the shunt capacitance and leakage between wires, are "distributed." Such networks are usually termed transmission lines (see also Section 14 in volume on Electric Power), but in the case of very high-frequency currents (5 meters or less) a network contained within an ordinary room may have to be similarly treated.

The usual transmission line in communication circuits consists of two similar parallel wires, or a single wire enclosed in a conducting cylindrical sheath. At every point on the line, some current flows from one wire to the other, or to the sheath, owing to capacitance and leakage conductance resulting from the imperfect dielectric between them. In consequence of this, the current in each conductor varies along the line. This is illustrated in Fig. 2, which shows an elementary section of a balanced line, or of a completely unbalanced

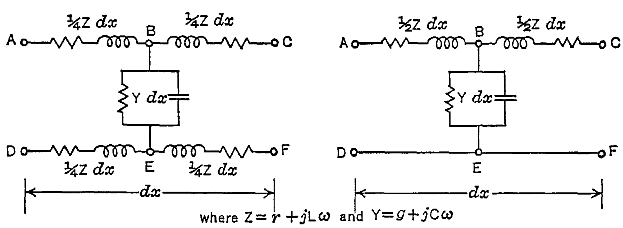


FIG. 2. Elementary Section of a Transmission Line

line equivalent to the coaxial type of construction. The current entering at A is not the same as that leaving at C, owing to the capacitance and conductance shunted between B and E. Hence the total drop due to the total resistance of the line is not this resistance multiplied by the current at any point on the line. Simple impedance equations cannot be written; simple differential equations can be given (for this solution see Section 14 in volume on Electric Power), and these offer one method of attack.

An alternative method is based on the assumption that such a uniform transmission line may be considered to be composed of an infinite number of symmetrical networks, each of which corresponds to an infinitesimal length of line. If r is the resistance and L the inductance per unit length (for two wires, sometimes called a loop-mile when the unit of length is a mile), and g the conductance and C the capacitance between the wires, or from the central conductor to the sheath, per unit of length, then, in a section of length dx, there will be a resistance $r\,dx$, an inductance $L\,dx$, conductance $g\,dx$, and capacitance $C\,dx$. As dx approaches zero the sections become smaller, and a line of these sections approaches a line with uniformly distributed constants.

VOLTAGE AND CURRENT RELATIONS. Let $Z = r + jL\omega = |Z| \underline{/\theta_z}$ be the series impedance per unit length, and let $Y = g + jC\omega = |Y| \underline{/\theta_y}$ be the shunt admittance per unit length of line. Then for the T network equivalent to an infinitesimal length of line (see Fig. 3) for which $\theta = d\theta$

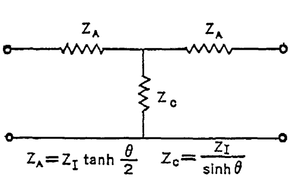


FIG. 3. T Network Equivalent to Smooth Line

$$Z_A = Z_I \frac{d\theta}{2} = \frac{Z}{2}\,dx \tag{8a}$$

$$Z_c = \frac{Z_I}{d\theta} = \frac{1}{Y\,dx} \tag{8b}$$

Thus

$$Z_I = \sqrt{\frac{Z}{Y}} = \sqrt{\frac{(r + jL\omega)}{(g + jC\omega)}} \tag{9a}$$

$$= \sqrt[4]{\frac{r^2 + L^2\omega^2}{g^2 + C^2\omega^2}} \Bigg/ \frac{1}{2}\left(\tan^{-1}\frac{L\omega}{r} - \tan^{-1}\frac{C\omega}{g}\right) \tag{9b}$$

This impedance is frequently called the *characteristic impedance* of the uniform line; the terms image impedance and iterative impedance which are also used are somewhat more clearly defined. They are, of course, identical to one another for the symmetrical line.

From eqs. (8) it follows that $d\theta = \sqrt{ZY}\,dx$, whence

$$\theta = \sqrt{ZY}\,x = \gamma x; \qquad \gamma = \sqrt{ZY} \tag{10}$$

where x represents the distance to a point on the line from the sending end. The quantity γ is called the *propagation constant* of the line and is evaluated *per unit length of line.* For a given length of line the total propagation constant is seen to be identical to the image transfer constant of the equivalent symmetrical network. γ is a complex number and can be expressed as $\gamma = \alpha + j\beta$, where α and β are real. Expanding eq. (10) and separating reals and imaginaries

$$\alpha = \sqrt{{}^1\!/_2[\sqrt{(r^2 + L^2\omega^2)(g^2 + C^2\omega^2)} + (rg - LC\omega^2)]} \tag{11a}$$

$$\beta = \sqrt{{}^1\!/_2[\sqrt{(r^2 + L^2\omega^2)(g^2 + C^2\omega^2)} - (rg - LC\omega^2)]} \tag{11b}$$

The parameters α and β are called the *attenuation constant* and *phase shift constant*, respectively, of the line. Note that they are both functions of frequency.

The relations given by eqs. (9) and (10) indicate the similarity between the behavior of networks having lumped constants and the behavior of uniform lines. At some point a distance x from the sending end of a line of length l the voltage E_x and current I_x are given by eqs. (4), (5), and (6) of article 1, where $\theta_1 = \gamma x$ and $\theta_2 = \gamma(l - x)$. Similarly the input impedance of the line is given by eq. (3a).

INCIDENT AND REFLECTED WAVES. The instantaneous voltage at any point on the wire is

$$e_x = \text{Real part } [E_1\epsilon^{j\omega t}(A_1\epsilon^{-\gamma x} + B_1\epsilon^{\gamma x})]$$

$$= |E_1|\,[|A_1|\,\epsilon^{-\alpha x}\cos(\omega t - \beta x + \delta_A) + |B_1|\,\epsilon^{\alpha x}\cos(\omega t + \beta x + \delta_B)] \tag{12}$$

and

$$i_x = \text{Real part}\left[\frac{E_1}{Z_I}\epsilon^{j\omega t}(A_1\epsilon^{-\gamma x} - B_1\epsilon^{\gamma x})\right]$$

$$= \left|\frac{E_1}{Z_I}\right|[|A_1|\,\epsilon^{-\alpha x}\cos(\omega t - \beta x + \delta_A - \phi) - |B_1|\,\epsilon^{\alpha x}\cos(\omega t + \beta x + \delta_B - \phi)] \tag{13}$$

where $A_1 = |A_1|\,\underline{/\delta_A}$, $B_1 = |B_1|\,\underline{/\delta_B}$, $Z_I = |Z_I|\,\underline{/\phi}$ and E_1 is real.

Each of these expressions is composed of two components, the *incident "wave"* which *decreases* in magnitude as x *increases*, and the *reflected "wave"* which *increases* as x *increases*. These components are called "waves" because they appear to travel along the wire with a velocity $v = \omega/\beta$. The distance between two consecutive points on the line at which $\cos(\omega t - \beta x + \delta_A)$ and its derivative have the same algebraic values for a fixed value of t is called a wavelength (λ), so that $\beta\lambda = 2\pi$. The time elapsing during a complete cycle of values is called the *period* (T), so that $T = 1/f$. In terms of these values the *velocity of phase propagation* of the waves is $v = \omega/\beta = \lambda/T$.

STANDING WAVES. The combination of the incident and reflected waves, adding sometimes in and sometimes out of phase, causes variation in the value of the voltage and current with position along the line. The expression for the amplitude of the voltage is

$$|E_x| = |E_1|\sqrt{|A_1|^2\epsilon^{-2\alpha x} + |B_1|^2\epsilon^{2\alpha x} + 2|A_1||B_1|\cos(2\beta x + \delta_B - \delta_A)} \tag{14}$$

and for the current,

$$|I_x| = \left|\frac{E_1}{Z_I}\right|\sqrt{|A_1|^2\epsilon^{-2\alpha x} + |B_1|^2\epsilon^{2\alpha x} - 2|A_1||B_1|\cos(2\beta x + \delta_B - \delta_A)} \tag{15}$$

each of which will have maximum and minimum values along the line whenever $|B_1| \neq 0$, since $\cos(2\beta x + \delta_B - \delta_A)$ changes more rapidly than $\epsilon^{2\alpha x}$. When attenuation is negligible and $|A_1| = |B_1|$

$$|E_x| = |E_1||A_1|\sqrt{2[1 + \cos(2\beta x + \delta_B - \delta_A)]} \tag{16}$$

From this it is evident that there are positions where $|E_x| = 2|E_1||A_1|$ called *voltage loops*, and others where $|E_x| = 0$ called *voltage nodes*. Similarly there are current loops and nodes, the current nodes corresponding to voltage loops, and vice versa.

It should be noted that the condition that $|A_1| = |B_1|$ requires that $\rho_2\epsilon^{-j2\beta l} = \pm 1.0$. Thus true *standing waves* occur only when the line is a multiple of a quarter wavelength and when $\rho_2 = \pm 1.0$, that is, when Z_2 is either zero or infinite.

INPUT IMPEDANCE. The input impedance of a uniform line terminated in an impedance Z_2 at a distance x from the sending end is given by

$$Z_i = Z_I \left[\frac{1 + \rho_2\epsilon^{-2\alpha x}\epsilon^{-j2\beta x}}{1 - \rho_2\epsilon^{-2\alpha x}\epsilon^{-j2\beta x}} \right] \tag{17}$$

where $\rho_2 = \dfrac{Z_2 - Z_I}{Z_2 + Z_I}$. If $Z_2 = Z_I$, then $\rho_2 = 0$ and $Z_i = Z_I$. The variation of Z_i with frequency is a smooth wave which approaches the value $\sqrt{L/C}$ as f increases.

When $Z_2 \neq Z_I$ the Z_i vs. f curve has maxima and minima which can be approximately located. If it be assumed that Z_I and Z_2 are pure resistances, which together with α are independent of frequency, then the locus of Z_i in the Z_i plane is a circle of radius

$$Z_I \left(\frac{2\rho_2\epsilon^{-2\alpha x}}{1 - \rho_2^2\epsilon^{-4\alpha x}} \right) \tag{18a}$$

with its center at

$$\left[Z_I \left(\frac{1 + \rho_2^2\epsilon^{-4\alpha x}}{1 - \rho_2^2\epsilon^{-4\alpha x}} \right), 0 \right] \tag{18b}$$

If ρ_2 is positive ($Z_2 > Z_I$), Z_i will have its maximum or minimum value when $\epsilon^{-j2\beta x}$ is $+1$ or -1. This requires that $2\beta x = 2n\pi$, or $(2n - 1)\pi$, where n is any integer. Since $\beta = \omega/v$ the frequency at which the maximum or minimum occurs is given by

$$f_{\text{max.}} = \frac{nv}{2x} \quad \text{or} \quad f_{\text{min.}} = \frac{(2n-1)v}{4x}$$

If $f = f_1$ when $n = n_1$ and $f = f_2$ when $n = n_1 + 1$, then

$$x = \frac{v}{2(f_2 - f_1)}$$

But since $f_2 - f_1$ is the frequency interval between two successive maxima, or two successive minima, the distance x to the point on the line at which there is an impedance discontinuity can be determined approximately. It should be noted that the expressions for the frequencies at which maxima and minima occur will be interchanged if ρ_2 is negative ($Z_2 < Z_I$).

DISTORTIONLESS LINES. In communication systems many frequency components are usually present, and it is desirable to have uniform attenuation with frequency. This condition exists on a line if $r/L = g/C$, in which case

$$\alpha = \sqrt{rg} \quad \text{and} \quad Z_I = \sqrt{\frac{L}{C}} = \sqrt{\frac{r}{g}}$$

both of which are independent of frequency. Similarly, $\beta = \omega\sqrt{LC}$, which is the condition for linear phase shift and no delay distortion. In this case also

$$v = \frac{\omega}{\beta} = \frac{1}{\sqrt{LC}}$$

which states that all waves are propagated with the same velocity.

TRANSIENTS IN NETWORKS

By Harold A. Wheeler

13. TRANSIENT DISTURBANCES

PROPERTIES OF TRANSIENTS. A transient disturbance, in its simplest concept, is one that occurs in a time interval separated from other disturbances. In general, a transient may be superimposed on other transients or continuous waves, according to the superposition theorem, while otherwise retaining its own characteristics. A single transient cannot be a periodic wave in the strict sense, although it may be a damped oscillation of a definite period. A borderline case is the "periodic transient," which is a periodic non-overlapping series of transients; each transient retains the properties of a transient while the series has the properties of a periodic wave.

Any periodic wave can be analyzed into a "Fourier series" which is a sum of sinusoidal components of frequencies in harmonic relation. The corresponding representation of a transient disturbance is possible by the "Fourier integral," which does not give a number of distinct components but rather the distribution of energy over the frequency spectrum. The Fourier series and the Fourier integral are analogous to the line spectrum and the band spectrum in light waves.

In general, an exact representation by a Fourier series requires an infinite number of components extending over the entire frequency spectrum, but a practical approximation requires only a limited number of components within a limited bandwidth. The same is true with respect to the limited bandwidth required for the practical reproduction of a transient. It is essential to consider the degree of approximation required or attained in any particular case, and to realize that both the duration of the transient and the frequency bandwidth are theoretically unlimited although practically limited by the sensitivity of the system.

Refer to Section 9, Pulse Techniques, for much information on transients as exemplified by pulses, with emphasis on their application in practical systems.

TYPES OF TRANSIENTS. It is convenient to define transient disturbances of several idealized types. An actual transient may be classified by its similarity to one of these types, or as the response of a certain network to one of these types.

The unit step is one of the elementary transients, shown in Fig. 1(a) and (b). It may be caused by switching or keying a current or voltage of unit amplitude. The form of Fig. 1(a) is also called the "Heaviside unit function," with a jump from zero to one. The form (b) has a jump from $-1/2$ to $+1/2$; it is preferred for some analytical purposes because it has zero d-c component. Campbell and Foster (references 7 and 8) designate the unit step as S_{-1}, called the singularity function of -1 order.

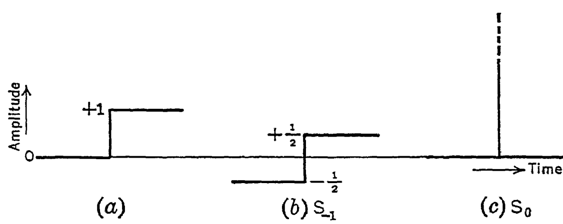


Fig. 1. Unit Step and Unit Impulse

The unit impulse, shown in Fig. 1(c), is unique among transients in that it has a uniform frequency spectrum. It is defined as the derivative or slope of the unit step; therefore it has unit area as the product of very large amplitude and very small duration. It is also called by Hansen the "delta function" (reference 38). Campbell and Foster (references 7 and 8) designate the unit impulse as S_0, called the singularity function of zero order.

In practice, the step is easier to generate because of its limited amplitude, and the transient response thereto has limited amplitude. The impulse may be approximated with limited amplitude if its duration is reduced until its frequency spectrum is substantially uniform over a frequency bandwidth sufficient for any particular tests. Most networks have some integrating action, so their response, even to an ideal impulse, would have limited amplitude.

Some oscillatory transients are illustrated in Fig. 2. In general, there are reversals of polarity, which may or may not be periodic in time. The transient response of a practical network is usually oscillatory in some degree. Figure 2(a) shows a damped oscillation which is a common occurrence; Fig. 2(b) shows a pulse-modulated wave including several cycles of a carrier wave, while (c) shows a carrier wave with a step in its modulation envelope. A carrier wave cannot be conceived as modulated by the ideal impulse, because the duration is insufficient to retain the identity of the carrier frequency.

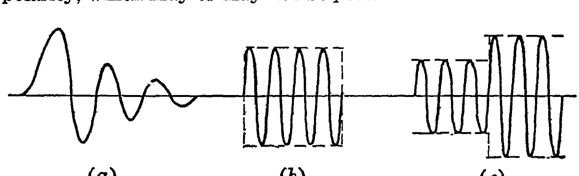


Fig. 2. Oscillatory Transients

Practical steps have finite slope, and practical pulses have finite duration and amplitude.

For test purposes, however, both can be made to approach the ideal as closely as required, so the transient response of a network to the ideal types can be tested.

A pulse to be used for modulating a carrier wave is sometimes denoted a "d-c" pulse to distinguish it from the modulated carrier of pulse envelope. The essential difference is in the frequency spectrum, the former having most of its energy concentrated in a band including zero frequency, and the latter in a band including the relatively high carrier frequency. The direct and carrier pulses are distinct concepts if their respective frequency bands are separate.

Amplitude modulation of a carrier wave generates symmetrical sidebands, and pulse modulation follows this rule. The carrier and both sidebands are needed for exact reproduction of the modulation. A small modulation superimposed on a continuous carrier, as illustrated in Fig. 2(*c*), can be approximately reproduced in a system responsive to one sideband and one-half the relative carrier amplitude, as employed in television (references 5, 11, 12, 16–19, 21, 24, 25, and Section 20).

FREQUENCY SPECTRUM. Figure 3 shows the frequency spectrum of several kinds of disturbances. The pure sine wave has a single frequency component (*a*). A periodic wave in general, such as a repeating short pulse, has a series of frequency components of relative amplitude indicated by an envelope of the lines (*b*). A transient has a continuous frequency spectrum (*c*). If the transient has the same shape as a single cycle of the periodic wave, the continuous spectrum (*c*) has the same frequency distribution as the envelope of lines (*b*).

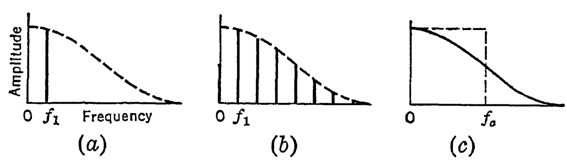


FIG. 3. The Frequency Spectrum of Periodic Waves and Aperiodic Disturbances

Periodic disturbances have a fundamental frequency, denoted f_1 in Fig. 3, and all components are harmonically related on integral multiples of this frequency. A transient lacks a fundamental frequency but, like the periodic wave, still requires frequency components over a certain band width for its reproduction with sufficient approximation. In Fig. 3(*c*), a nominal cutoff frequency f_c may be defined in such a way as to include most of the energy of the frequency spectrum. For pulse operation, the nominal cutoff frequency on the frequency spectrum of amplitude may be defined as the boundary of a rectangle (drawn as shown in dotted lines) having the same area as the amplitude spectrum (reference 11).

In the line spectrum of a periodic wave, each component has a definite amplitude which may be expressed in terms of current or voltage (or analogous linear quantities). In the band spectrum of a transient, however, the amplitude is expressed per unit of frequency bandwidth, in units such as the volt per cycle per second or volt-second. This concept is elusive, but the shape of the spectrum indicates clearly the relative importance of the various frequency components.

As an alternative, the frequency spectrum may be presented in terms of relative energy instead of amplitude. It is then expressed in terms of energy per unit of frequency bandwidth, in units such as the joule per cycle per second or joule-second. The area under the curve is the total energy of the transient.

The frequency spectrum of Fig. 3 shows the relative response of a receiver of very narrow and constant bandwidth as it is tuned over the frequency range. Spectrum analyzers are operated on this principle, some of which show the spectrum directly on an oscilloscope (Section 9).

SPEED OF INFORMATION. In communication of any kind, the available frequency bandwidth is limited, and this may restrict the speed of transmission (references 2, 4–6, 9–11, and Section 9). This is one of the principal limitations in picture transmission.

As a simple example, Fig. 4 shows the code pulses for the word "as." The pulse pattern (*a*) contains dots and dashes in the form of short and long pulses. It is necessary to distinguish the presence or absence of a pulse at intervals of one pulse width.

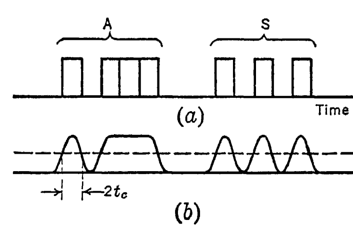


FIG. 4. Speed of Information by Pulses

Increasing the speed or decreasing the frequency bandwidth up to a certain limit has the effect of rounding the pulses (*b*) but not filling in the space between pulses. This much distortion is permissible and economical in practice.

The rounded pulses in Fig. 4(*b*) have a frequency spectrum similar to that in Fig. (3*c*) if the dots have a width $2t_c$ related to the nominal cutoff frequency f_c as follows (reference 11):

$$2t_c = \frac{1}{2f_c} \tag{1}$$

This means that the shortest pulse or space has a duration of 1/2 cycle at the nominal cutoff frequency. The speed of information is proportional to the number of pulses in a given interval which is proportional to the frequency bandwidth.

A completely modulated carrier requires twice the bandwidth because double-sideband operation is essential. A partially modulated carrier, with single-sideband operation as in television, requires only slightly more bandwidth than the modulating pulses but also requires more power to surmount background noise and interference.

14. BEHAVIOR OF NETWORKS

DIFFERENTIATION AND INTEGRATION. A linear network operates only on the amplitude and phase of a sine wave, retaining the wave form. Though the same is true of each component of a transient, the operation on all components may greatly change its shape. The simplest distorting operations of networks are differentiation and integration.

The two basic differentiating networks are shown in Fig. 5. In each case, the input and output are denoted by subscripts 1 and 2. In (*a*) and (*b*), the series capacitor C and the shunt inductor L are so connected that each gives an output proportional to the time derivative of the input;

$$\text{Fig. 5}(a) \quad \frac{dE_1}{dt} C = I_2; \qquad (b) \quad \frac{dI_1}{dt} L = E_2 \tag{2}$$

The instantaneous voltage and current are here denoted E and I.

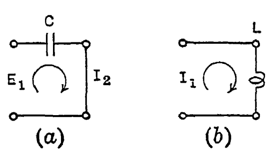


FIG. 5. Basic Differentiating Networks

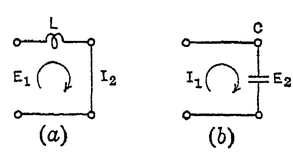


FIG. 6. Basic Integrating Networks

If instead the amplitudes of the frequency components are denoted E and I, the corresponding relations for any particular frequency may be written:

$$\text{Fig. 5}(a) \quad E_1 j\omega C = I_2; \qquad (b) \quad I_1 j\omega L = E_2 \tag{3}$$

Here j is the quadrature factor $\sqrt{-1}$ and $\omega = 2\pi f$ is the radian frequency. The two sets of equations are alike in form except that d/dt is replaced by $j\omega$; this explains why $j\omega$ is sometimes called a differential operator. It includes the inseparable two essentials of differentiation, namely, an amplitude ratio directly proportional to frequency and a leading phase shift of one quadrant.

An example of differentiation is the conversion of a unit step to a unit impulse. The former has frequency components of amplitude inversely proportional to frequency, while differentiation changes it to the impulse having frequency components of uniform amplitude.

The two basic integrating networks are shown in Fig. 6. In (*a*) and (*b*), the series inductor L and the shunt capacitor C are so connected that each gives an output proportional to the time integral of the input:

$$\text{Fig. 6}(a) \quad \left(\int E_1\,dt\right)\frac{1}{L} = I_2; \qquad (b) \quad \left(\int I_1\,dt\right)\frac{1}{C} = E_2 \tag{4}$$

Changing the significance of voltage and current, E and I, from the instantaneous values above to the amplitudes of the frequency components:

$$\text{Fig. 6}(a) \quad E_1 \frac{1}{j\omega L} = I_2; \qquad (b) \quad I_1 \frac{1}{j\omega C} = E_2 \tag{5}$$

Since $\int dt$ is replaced by $1/j\omega$, the latter is sometimes called an integral operator. It includes the inseparable two essentials of integration, namely, an amplitude ratio inversely proportional to frequency and a lagging phase shift of one quadrant. An example of integration is the conversion of a unit impulse to a unit step.

The basic networks of Figs. 5 and 6 rely on simple admittance or impedance coupling, so the input and output are voltage and current in one order or the other. Approximate differentiation and integration can be obtained by voltage-ratio or current-ratio coupling, the former being shown in Figs. 7 and 8. Each of these networks includes a resistor R in

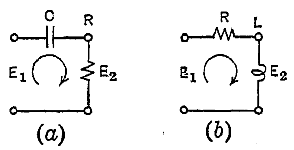


FIG. 7. Voltage-ratio Differentiating Networks

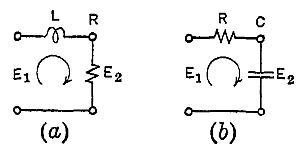


FIG. 8. Voltage-ratio Integrating Networks

addition to a reactor C or L. The required approximation to the ideal operation in each case places certain requirements on the time constant of the network, which is L/R or CR.

Figure 9 shows the meaning of the time constant in a charging or discharging operation (*a*) or (*b*). In each case the transient has an exponential variation with time, and the time constant t_1 is the length of time required to approach completion of the operation. Quantitatively, it is the time to go to $(1 - 1/\epsilon)$ or 0.63 of completion ($\epsilon = 2.72$, the base of natural logarithms).

Though the concept of charging and discharging is commonly associated with energy storage by the voltage on a capacitor, it is equally applicable to energy storage by the current in an inductor.

A charging operation shown in Fig. 9(*a*) is exemplified by a voltage step applied to an integrating network of Fig. 8. The integrating operation continues only for a duration less than the time constant, so the time constant must be longer than the required period of approximate integration. This is a general rule for such integrating networks.

A discharging operation shown in Fig. 9(*b*) is exemplified by a voltage step applied to a differentiating network of Fig. 7. The differentiating operation is prolonged for a duration exceeding the time constant, so the time constant must be shorter than the permissible duration. This is a general rule for such differentiating networks.

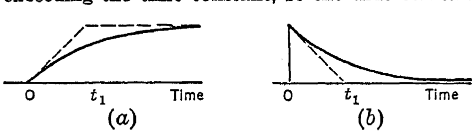


FIG. 9. Time Constants of Charging and Discharging

Meeting these conditions in Figs. 7 and 8 is promoted by a large value of series resistance or a small value of shunt resistance, so the voltage ratio is small for the frequency components of major importance.

Figure 10 is a chart of the charging and discharging operations obtained from the networks of Figs. 7 and 8 in response to steps and impulses.

OSCILLATIONS. If differentiation and integration are mixed in a network by combining CLR, the result is resonance. The transient response to a step or impulse may then be a damped oscillation. Figure 11 shows examples of voltage-ratio resonant networks and the response of each to certain input transients. A resonant network is one having maximum response at the frequency of resonance; an antiresonant network is one having minimum response at that frequency. Either one exhibits damped oscillations after a transient disturbance. The time constant of damping in the series-resonant circuit is $2L/R$, while that in the parallel-resonant circuit (with parallel R) is $2CR$.

REPEATING NETWORKS. There are many kinds of networks which respond to a step or impulse with interesting and significant output transients. One of the simplest is the integrating network of Fig. 8(*b*) repeated in successive stages of a vacuum-tube amplifier (No. 524.2 in references 7 and 8, also reference 38). Figure 12 shows the response of n such networks to an impulse. The integrating action, denoted by the time constant t_1, both delays and widens the pulse by virtue of the energy storage in each capacitor and its subsequent discharge (by repeater action) into the next capacitor. The delay of the pulse exceeds the widening, so this system is a crude delay network. The pulse peak is delayed by $(n - 1)t_1$. For large values of n it is widened to $\sqrt{2\pi}\sqrt{n-1}\,t_1$, and it approaches the symmetrical shape of a probability curve.

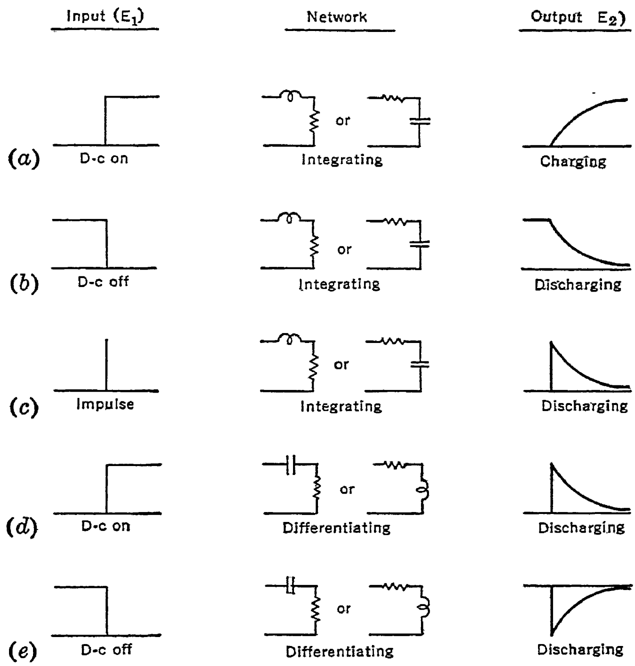


FIG. 10. Transient Response of Simple Networks

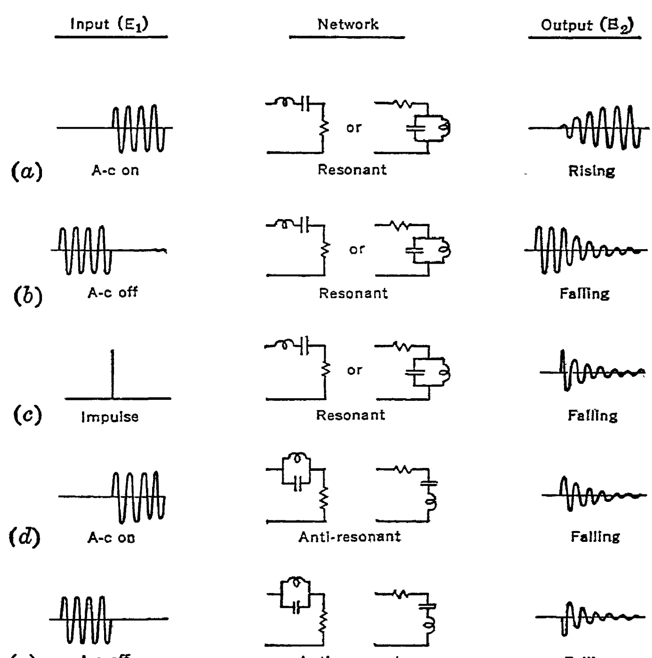


FIG. 11. Transient Response of Oscillatory Networks

The response of a resonant circuit to an impulse, as exemplified in Fig. 11(c) for a single network, may be extended to repeating networks. The envelope of the resulting transient oscillation then assumes the form of Fig. 12.

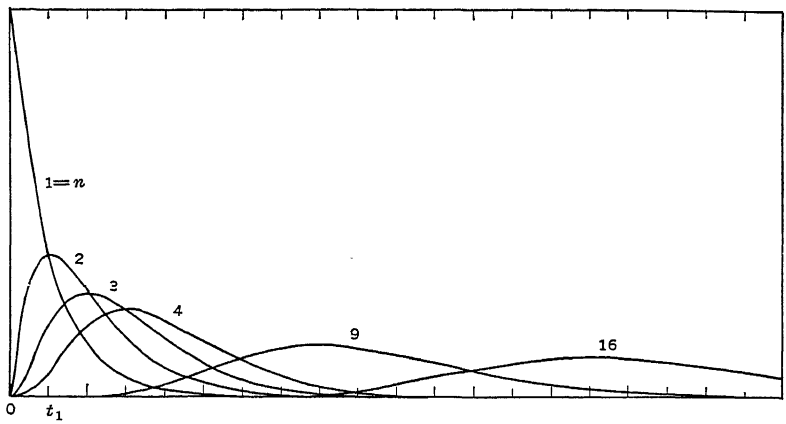


FIG. 12. The Delay and Widening of a Pulse by Repeating Integrating Networks

BANDWIDTH. The integrating action of a shunt capacitor limits the speed of information by restricting the frequency bandwidth. This is a major factor in a wide-band amplifier for such uses as television and radar, because each interstage coupling from one tube to the next has inherent shunt capacitance. Figure 13 shows a resistance-coupled amplifier stage subject to inherent shunt capacitance C across the coupling resistor R.

The frequency variation of the response of such an amplifier is shown in Fig. 14, in which f_1 is the frequency at which the reactance of the capacitor ($1/[2\pi fC]$) is equal to the shunt resistance (R). Figure 14(a) shows the amplitude variation for several conditions, starting with (1) simply R and C. The bandwidth is increased (2) by adding inductance L to build up the impedance by a tendency to resonance. A more complicated network,

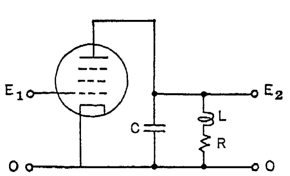


FIG. 13. Amplifier with Shunt Capacitance Limiting the Bandwidth

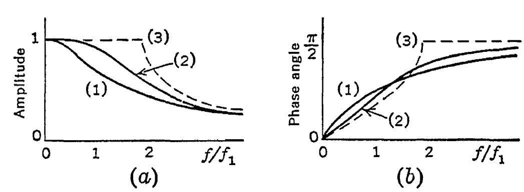


FIG. 14. The Limitation of Bandwidth by Shunt Capacitance

termed the "dead-end filter" (reference 15), can be used to extend the bandwidth as far as curve (3) but no further. The maximum bandwidth over which a uniform amplitude ratio can be obtained is theoretically

$$2f_1 = \frac{1}{\pi CR} \tag{6}$$

In practice, about half this bandwidth is obtained in simple circuits with a sufficient approximation to uniformity.

If many stages of wide-band amplification are needed, the phase distortion shown in Fig. 14(b) may be a limiting factor more severe than amplitude distortion. The ideal phase variation is a linear proportionality to the frequency. Condition (1), with simply R and C, yields a convex phase curvature, but in this case the amplitude distortion is more severe. (See Fig. 12, which applies to this case.) Condition (2), with L added, yields a concave phase curvature which happens to be more detrimental than the residual amplitude distortion. There are cases (notably in radar pulse receivers) where a compromise between (1) and (2) may be optimum (reference 38). The extreme condition (3) yields abrupt changes in amplitude and phase at cutoff, which cause transient damped oscillations.

AMPLITUDE AND PHASE DISTORTION. The transient response of linear networks is related uniquely with their steady-state characteristics, so a knowledge of the latter makes it possible to estimate the transient response. This is the province of the Fourier integral and other concepts such as "paired echoes" for evaluating distortion (references 14 and 20). Examples of amplitude and phase distortion are shown in Fig. 14 for the simple case of Fig. 13 described above.

Amplitude distortion may be simply the limitation of the bandwidth or may also include irregularities within the bandwidth. It is generally expressed in terms of attenuation, since this is a logarithmic quantity which can simply be added for cumulative stages (p. 1-37).

Phase distortion is any departure from linear proportionality between the lagging phase angle and the frequency. It is expressed in angular measure, which is additive for cumulative stages. The absolute unit of angle is the radian, which is $1/2\pi$ circle or 57.3°. The radian is quantitatively comparable with the napier, so 6.6° is comparable with 1 db. The inherent properties of passive linear networks determine certain relations between attenuation and phase angle (references 36 and 41). For any pattern of variation of the attenuation over the entire frequency range, there is a corresponding pattern of minimum phase angle obtainable in networks. Those networks which provide selective attenuation with minimum phase angle are termed "minimum-phase" networks and others "excess-phase" networks. Minimum-phase networks include self-impedance couplings, simple ladder networks, and any other network whose response can be expressed as a product of physically realizable self-impedance factors. Excess-phase networks include transmission lines, all-pass phase-correcting networks (lattice or bridged-tee), and networks with negative mutual inductance.

Phase-correcting networks are theoretically possible free of attenuation, but not attenuation-correcting networks free of phase distortion. Therefore it is customary, in the design of a practical network, first to obtain the required attenuation, and then if necessary to add phase-correcting networks for obtaining linear phase. Usually the phase correction need be effective only over the band width of nearly maximum response. Phase correction (free of attenuation) always increases the phase angle, never the reverse.

Amplitude distortion, free of phase distortion, cannot destroy the symmetry of a symmetrical input pulse. Therefore any asymmetrical distortion is a symptom of departure from phase linearity.

A sharp cutoff at the edge of the useful bandwidth causes a damped oscillation or "overshoot" in the transient response to an impulse or, in less degree, to a step. If this result is not permissible, a gradual cutoff is required, as shown in Fig. 3(*c*) (reference 11).

15. GENERAL PRINCIPLES

THE FOURIER INTEGRAL. In the study and design of networks to handle transient disturbances, the most powerful concept is the Fourier integral. It is an extension of the more familiar Fourier series, which is restricted to periodic waves but still serves as an introduction to the integral. Each is essentially a relationship between a disturbance over a period of time and its frequency components; or, conversely, a set of components can be synthesized into the form of disturbance.

The following presentation of the Fourier series is in a form well adapted for extension to the integral and useful for direct application.

A wave form (of voltage or current, for example) is denoted $T(t)$ and is completed in the time interval between $-t_1/2$ and $+t_1/2$. The same wave form is repeated in successive intervals of the same period t_1, as a periodic wave. This wave can be expressed as a sum of sine-wave components of harmonic frequencies and the proper phase:

$$T(t) = \sum_{n=-\infty}^{\infty} F_n \exp\,(j2\pi n f_1 t) \tag{7}$$

The fundamental frequency is $f_1 = 1/t_1$. Each harmonic component has a frequency nf_1 which is an integral multiple of the fundamental frequency. Its amplitude is F_n, which is generally complex to include the phase angle. Obviously the dimensional units of $T(t)$ and F_n are the same.

The sine-wave nature of each component is indicated by the unit vector rotating at a frequency nf_1:

$$\exp\,(j2\pi n f_1 t) = \text{cis}\,(2\pi n f_1 t) = \cos\,(2\pi n f_1 t) + j \sin\,(2\pi n f_1 t) \tag{8}$$

Since $T(t)$ is real, it is apparent that the imaginary parts of each component in the summation must cancel out. This cancellation occurs between the amplitudes F_n of the components of equal positive and negative values of n. The zero-frequency (direct) component ($n = 0$) always has a real amplitude F_0.

The (complex) amplitude of each component is formulated as

$$F_n = \int_{-t_1/2}^{t_1/2} T(t) \exp(-j2\pi n f_1 t)\, dt \tag{9}$$

This integral selects and evaluates each harmonic of frequency nf_1. A tabular or graphical or formal integration can be used to compute F_n, which in general will be complex.

The extension of the series to the integral requires two more concepts. First, since a transient is presumed to occur only once, the period t_1 is made very large and the corresponding fundamental frequency f_1 very small. Secondly, the harmonic amplitude F_n is changed to the frequency spectrum $F(f)$ which represents the "amplitude-frequency density" or "amplitude per unit of frequency" in the vicinity of the frequency f (instead of nf_1). The density does not have the same dimensional units as $T(t)$ but rather the same units multiplied by "time." For example, if $T(t)$ is in volts, $F(f)$ is in volts per cycle per second, or volt-seconds.

With these steps, there follows the Fourier integral for expressing the wave form of a transient in terms of its frequency spectrum $F(f)$:

$$T(t) = \int_{-\infty}^{\infty} F(f) \exp(j2\pi ft)\, df \tag{10}$$

The other form expresses the frequency spectrum in terms of the transient:

$$F(f) = \int_{-\infty}^{\infty} T(t) \exp(-j2\pi ft)\, dt \tag{11}$$

Following Campbell and Foster (references 7 and 8), the above forms are symmetrical in that $F(f)$ and $T(t)$ are interchangeable merely by reversing the sign of j. This is accomplished by integrating from $-\infty$ to $+\infty$, and expressing in terms of the variables f (instead of $\omega = 2\pi f$) and t, which are mutually reciprocal. These forms include $j2\pi$ under the exponential function, which is logical since $j2\pi$ is the natural logarithm of a unit vector rotated by one cycle, and $\exp(j2\pi)$ symbolizes that vector.

In electrical networks, the interchangeability of $F(f)$ and $T(t)$ has reality only if F is real, since T is real. This is true only in idealized networks, which may be instructive examples. In such cases, both F and T are symmetrical in form.

The direct significance of $F(f)$ and $T(t)$ is simple. It is based on the unique fact that an impulse has a uniform amplitude density over the spectrum. If a unit impulse is applied to a network which modifies its frequency spectrum to the form $F(f)$, the output is a transient of the form $T(t)$.

If a transient T_1 of any shape is applied to any network, the output transient T can be expressed and often evaluated simply by the following procedure: (1) compute F_1, the spectrum of the input transient T_1; (2) formulate F_2, the frequency response of the network; (3) note that the frequency spectrum of the output transient is $F = F_1F_2$; (4) from the knowledge of F, use the Fourier integral to express the output transient T. If the input is a unit step instead of a unit impulse, $F_1 = 1/j2\pi f$.

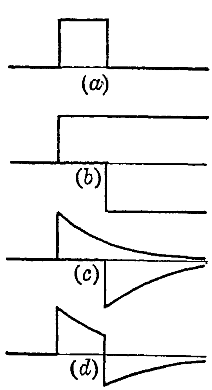


FIG. 15. An Example of Superposition

Closely related methods of transient analysis and synthesis are the Laplace transformation and the Heaviside operational calculus. The Fourier integral is a restricted case of the Laplace transform, but actually the one which is most simply adapted to the study of transients in electrical networks. The operational calculus is merely a process for deriving, explaining, and applying some of the ideas inherent in the Fourier integral. The present state of the literature places the Fourier integral in a position to be of the greatest aid in solving network problems.

THE SUPERPOSITION THEOREM. Any transient can be regarded as built up of a number of steps or pulses, a large number for smooth wave forms. This is the basis of the superposition theorem. A simple example, and probably the first to be discovered, is shown in Fig. 15. A square pulse (*a*) of any width is the resultant of two superimposed steps (*b*), the first one positive and the second negative. The transient response of a network to the square pulse is obtained by first evaluating its response to each step (*c*) and superimposing these transients (*d*).

The F_1F_2 procedure described above has an alternative in the superposition theorem of the operational calculus. This theorem is encompassed in a single Fourier integral (No. 202 in references 7 and 8). As above, T_1 is the input transient (whose frequency spectrum is F_1) and F_2 is the response of the network; T_2 is the transient which would be obtained from the network in response to a unit impulse. The output transient is then expressed in terms of T_1 and T_2 as follows:

$$T(t) = \int_{-\infty}^{\infty} F_1(f)F_2(f) \exp(j2\pi ft)\, df \tag{12}$$

$$= \int_{-\infty}^{\infty} T_1(t')T_2(t - t')\, dt' = \int_{-\infty}^{\infty} T_1(t - t')T_2(t')\, dt' \tag{13}$$

The symbol t' denotes the variable of integration, as distinct from t, the time variable in the transient.

THE ENERGY INTEGRAL. One of the most useful corollaries of the Fourier integral is the "energy integral":

$$\int_{-\infty}^{\infty} |F|^2\, df = \int_{-\infty}^{\infty} T^2\, dt \tag{14}$$

It states that the energy of the transient is proportional to the area under the energy-distribution curve over the frequency range, this curve being plotted in terms of $|F|^2$. This concept has the widest use for the evaluation of the power associated with random noise, which behaves as a great number of impulses occurring at random. It is noted that the total energy is determined by the amplitude (squared) independent of phase.

IDEALIZED FILTERS. Idealized examples of filters and transients are instructive as to the basic limitations. The most common such example is shown in Fig. 16 (references

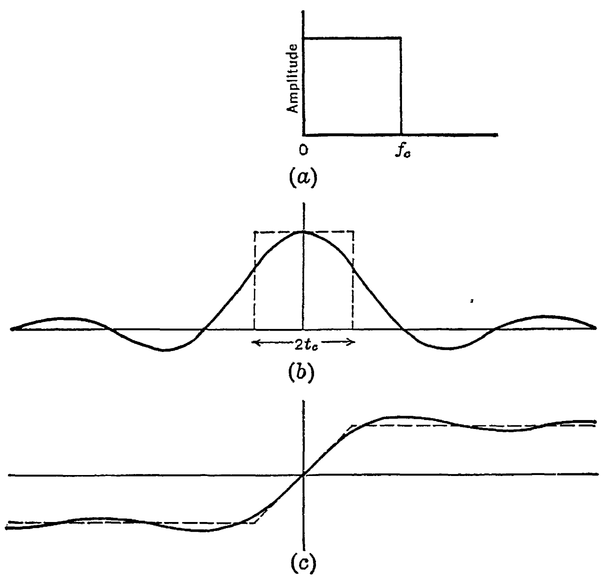


FIG. 16. An Idealized Filter and Its Transient Response

9 and 10). A network (*a*) has uniform response (with zero phase angle) up to a cutoff frequency f_c, and no response above this frequency. If an impulse is applied to this network, the output transient (*b*) is symmetrical and is accompanied by transient oscillations at the cutoff frequency. If this response is approximated in a real network, the phase slope is rather great and delays the main pulse so much that the earliest perceptible oscillations (preceding the pulse) occur at a time later than the applied impulse. Incidentally, this requires an "excess-phase" network as defined above.

In the transient of Fig. 16(*b*), the nominal pulse duration may be defined as $2t_c$, determined by the dotted rectangle of equal net area. If a step (instead of an impulse) is applied to the same network, the output (*c*) has a slope of the same nominal duration (called the "slope time" or "time of rise" or "build-up time").

The ideal filter of Fig. 16(*a*) has a perfectly sharp cutoff, which causes transient oscillations in the output (*b*) and (*c*). The opposite extreme is a response having gradual cutoff of the form of a probability curve, somewhat similar to Fig. 3(*c*) (reference 11). In response to an impulse, such a filter yields a rounded symmetrical pulse of the same shape, free of overshoot and transient oscillations. This performance is approximated by a large number of networks like Fig. 8, with output transients shown in Fig. 12. This case has been found ideal in radar receivers, to assure the prompt damping of one echo pulse in readiness for another.

DELAY. A lagging phase angle is characteristic of filters passing a limited band width (low-pass and band-pass), and also of transmission lines. The result is a delay of the signal, and perhaps also a distortion of its wave form.

The delay is defined in different ways, as a function of frequency. If β is the lagging phase angle (in radians) at a frequency ω (in radians per second), the "intercept delay" or "phase delay" (in seconds) is merely β/ω. It signifies the delay of each sine-wave component along the time axis. If the delay is the same for all components, it is free of distortion. This result requires "linear phase," that is, a phase angle directly proportional to the frequency.

The more general concept is the "envelope delay," defined as the "phase slope" $d\beta/d\omega$. This definition is free of the ambiguity of multiples of 2π in determining the phase angle. It derives its name from its significance as the delay of the envelope of a transient oscillation of many cycles, as distinguished from the delay of the individual cycles. The delay of the cycles is inconsequential if the transient is to be rectified, in which event the envelope delay uniquely determines the effect of the phase angle on the output of the rectifier.

These two definitions of delay correspond with those of the wave velocity in a transmission line or other wave medium. The intercept delay or phase delay determines the "phase velocity" or "steady-state velocity," with its ambiguities. The envelope delay uniquely determines the "group velocity," which, from one point to another, cannot exceed the speed of light.

BIBLIOGRAPHY

1. C. P. Steinmetz, *Transient Electric Phenomena and Oscillations*, third ed. McGraw-Hill (1920). (Many original concepts, mainly superseded by later methods of presentation.)
2. R. V. L. Hartley, Relations of Carrier and Sidebands in Radio Transmission, *Proc. I.R.E.*, Vol. 11, 34–56 (February 1923). (Double or single sideband, with or without carrier.)
3. L. A. Hazeltine, *Electrical Engineering*. The Macmillan Co. (1924). Chapter 6, "Transient Currents and Electric Waves," pp. 210–254. (The transient response in simple circuits and lines.)
4. H. Nyquist, Certain Factors Affecting Telegraph Speed, *B.S.T.J.*, Vol. 3, 324–352 (April 1924). (The limitation imposed by the frequency band width and the choice of a code.)
5. H. Nyquist, Certain Topics in Telegraph Transmission Theory, *Trans. A.I.E.E.*, Vol. 47, 617–644 (April 1928). (The relation between speed and frequency bandwidth; the case of single sideband with half-amplitude carrier.)
6. R. V. L. Hartley, Transmission of Information, *B.S.T.J.*, Vol. 7, 535–563 (July 1928). (The amount of information proportional to the product of the frequency bandwidth and the time.)
7. G. A. Campbell, Practical Application of the Fourier Integral, *B.S.T.J.*, Vol. 7, 639–707 (October 1928). (Purely mathematical introduction; table of paired coefficients.)
8. G. A. Campbell and R. M. Foster, Fourier Integrals for Practical Applications, *Bell Tel. Sys. Tech. Publ.*, Monograph B-584, September 1931. (Purely mathematical introduction; the most extensive tables of paired coefficients.)
9. A. T. Starr, Transients in Networks, *Electric Circuits and Wave Filters*, Sir Isaac Pitman & Sons, Ltd., Chapter 12, pp. 332–353, 1934. (Introduction to the concept of idealized networks and their transient response.)
10. E. A. Guillemin, *Communication Networks*, Vol. II. John Wiley (1935). Chapter 11, "The Transient Behavior of Filters," pp. 461–507.
11. H. A. Wheeler and A. V. Loughren, The Fine Structure of Television Images, *Proc. I.R.E.*, Vol. 26, 540–575 (May 1938). (The band-width limitation, especially the effect of sharp cutoff or gradual cutoff; a simple introduction to the Fourier integral in the absence of phase distortion; bibliography.)
12. J. E. Smith, B. Trevor, and P. S. Carter, Selective Sideband vs. Double Sideband Transmission of Telegraph and Facsimile Signals, *R.C.A. Rev.*, Vol. 3, 213–238 (October 1938).
13. A. V. Bedford and G. L. Fredendall, Transient Response of Multistage Video-frequency Amplifiers, *Proc. I.R.E.*, Vol. 27, 277–284 (April 1939).
14. H. A. Wheeler, The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *Proc. I.R.E.*, Vol. 27, 359–385 (June 1939).
15. H. A. Wheeler, Wide-band Amplifiers for Television, *Proc. I.R.E.*, Vol. 27, 429–438 (July 1939). (Fundamental limitations imposed by shunt capacitance; idealized networks including dead-end filters.)
16. R. Urtel, Observations Regarding Television Transmission by Single Sideband, *Telefunken Hausmitteilungen*, No. 81, pp. 80–83 (July 1939). (The envelope distortion in the transmission of single sideband and half-amplitude carrier through a filter of idealized characteristics.)
17. Leon S. Nergaard, A Theoretical Analysis of Single-sideband Operation of Television Transmitters, *Proc. I.R.E.*, Vol. 27, 666–677 (October 1939).
18. S. Goldman, Television Detail and Selective-sideband Transmission, *Proc. I.R.E.*, Vol. 27, 725–732 (November 1939). (The envelope distortion in the transmission of single sideband and half-amplitude carrier through a filter of idealized characteristics.)
19. H. Nyquist and K. W. Pfleger, Effect of the Quadrature Component in Single-Sideband Transmission, *B.S.T.J.*, Vol. 19, 63–73 (January 1940). (The envelope distortion in the transmission of a single sideband and half-amplitude carrier through a filter of idealized characteristics.)

20. Von F. Strecker, The Influence of Small Phase Distortion on the Reproduction of Television Signals, *E.N.T.*, Vol. 17, 51–56 (March 1940); The Effect of Amplitude and Phase Distortion, 93–107 (May 1940). (Paired echoes.)
21. R. D. Kell and G. L. Fredendall, Selective Sideband Transmission in Television, *R.C.A. Rev.*, Vol. 4, 425–440 (April 1940). (Response of single-sideband receiver to a unit step of modulation.)
22. M. I. T. Engineering Staff, *Electric Circuits.* John Wiley (1940). (A comprehensive treatment of networks with emphasis on their transient response.)
23. H. E. Kallmann, Transversal Filters, *Proc. I.R.E.*, Vol. 28, 302–310 (July 1940). (The synthesis, by a pattern of echoes, of the signal which would be produced by an idealized filter free of phase distortion.)
24. H. E. Kallmann and R. E. Spencer, Transient Response of Single-sideband Systems, *Proc. I.R.E.*, Vol. 28, 557–561 (December 1940). (The envelope distortion in the transmission of single sideband and half-amplitude carrier through a filter of idealized characteristics.)
25. C. P. Singer, A Mathematical Appendix to Transient Response of Single-sideband Systems, *Proc. I.R.E.*, Vol. 28, 561–563 (December 1940).
26. H. S. Carslaw and J. C. Jaeger, *Operational Methods in Applied Mathematics.* Oxford University Press (1941). (The Laplace transformation.)
27. H. A. Wheeler, The Solution of Unsymmetrical-sideband Problems with the Aid of the Zero-frequency Carrier, *Proc. I.R.E.*, Vol. 29, 446–458 (August 1941).
28. H. A. Wheeler, Common-channel Interference between Two Frequency-modulated Signals, *Proc. I.R.E.*, Vol. 30, 34–50 (January 1942). (The response to frequency modulation, by an analysis which is general for any wave form of modulation, either transient or periodic.)
29. J. B. Russell, Heaviside's Direct Operational Calculus, *Elec. Eng.*, Vol. 61, 84–88 (February 1942). (A brief introduction.)
30. R. V. L. Hartley, A More Symmetrical Fourier Analysis Applied to Transmission Problems, *Proc. I.R.E.*, Vol. 30, 144–150 (March 1942).
31. W. L. Sullivan, Fourier Integrals, *Elec. Eng.*, Vol. 61, 248–256 (May 1942). (An excellent introduction and bibliography.)
32. H. Salinger, Transients in Frequency Modulation, *Proc. I.R.E.*, Vol. 30, 378–383 (August 1942.)
33. A. V. Bedford and G. L. Fredendall, Analysis, Synthesis, and Evaluation of the Transient Response of Television Apparatus, *Proc. I.R.E.*, Vol. 30, 440–457 (October 1942).
34. M. F. Gardner and J. L. Barnes, *Transients in Linear Systems*, Vol. 1, "Lumped-constant Systems." John Wiley (1942). (The most intensive treatment of the subject, based on the Laplace transformation, with emphasis on electromechanical analogues. A complete bibliography.)
35. J. G. Brainerd, *Ultra-High-Frequency Techniques.* D. Van Nostrand (1942). Chapter 1, "Linear Circuit Analysis," pp. 1–47. (Transient response to steps and pulses.)
36. F. E. Terman, Relation between Attenuation and Phase Shift in Four-terminal Networks, *Proc. I.R.E.*, Vol. 31, 233–240 (May 1943).
37. Harold Pender and S. R. Warren, *Electric Circuits and Fields.* McGraw-Hill (1943). Chapter 4, "Introduction to Transient Phenomena in Electric Circuits," pp. 108–133.
38. W. W. Hansen, Transient Response of Wide-band Amplifiers, *Proc. of National Electronics Conference*, Vol. 1, 544–554 (1945); also *Electronic Industries*, Vol. III, No. 11, 80–82, 218, 220 (November 1944). (A definition of bandwidth for transients, and its application to amplifiers of one to 100 stages.)
39. R. V. Churchill, *Modern Operational Mathematics in Engineering.* McGraw-Hill (1944).
40. W. W. Hansen, On Maximum Gain-bandwidth Product in Amplifiers, *J. Applied Phys.*, Vol. 16, 528–534 (September 1945).
41. H. W. Bode, *Network Analysis and Feedback Amplifier Design.* D. Van Nostrand (1945). (A comprehensive treatment with emphasis on stability and bandwidth.)
42. H. E. Kallmann, R. E. Spencer, and C. P. Singer, Transient Response, *Proc. I.R.E.*, Vol. 33, 169–195 (March 1945).
43. M. Levy, The Impulse Response of Electrical Networks, *Elec. Com.*, Vol. 22, No. 1, 40–55 (1944). (Idealized networks.)
44. G. W. Carter, *The Simple Calculation of Electrical Transients.* The Macmillan Co. (1945). (A short mathematical course based on operational methods, with well-selected tables and bibliography.)
45. Ernest Frank, *Pulsed Linear Networks.* McGraw-Hill (1945). (An introductory course limited to the direct application of differential equations.)
46. D. G. Tucker, Transient Response of Filters, *Wireless Engineer*, Vol. XXIII, No. 269, pp. 36–42 (February 1946); Vol. XXIII, No. 270, 84–90 (March 1946).

Note: Reference 31 is recommended as a good introduction to the more advanced concepts, and a good bibliography to that date. References 29 and 31, with three related papers, have been reprinted by A.I.E.E. under the title *Advanced Methods of Mathematical Analysis* (1942).

NON-LINEAR ELECTRIC CIRCUITS

By Knox McIlwain

In solving the fundamental electric circuit equation

$$e = ri + L\frac{di}{dt} + \frac{1}{C}\int i\,dt \tag{1}$$

the assumption is often made that r, L, and C are constants. Circuits in which these conditions exist are called "linear circuits" and have been treated above. If any one or more of these circuit parameters vary with the current through it or the voltage across it, the solution of eq. (1) in general takes the form of an infinite series, so that the current wave form is not a replica of the voltage wave form. Such circuits are called "non-linear circuits."

NON-LINEAR DISTORTION. When it is desired that the wave form of the current through a particular circuit element be the same as the wave form of the original voltage, and when the response of the element is not directly proportional to the driving force (non-linear circuit), the element is said to introduce *non-linear distortion.* When the input voltage is a simple sine wave, the output current will contain components of double and higher multiples of the impressed frequency, the amplitudes being determined by the series solution mentioned above.

Non-linear distortion is most serious in sound reproduction since the spurious sound harmonics are readily noted by the ear and produce a very unpleasant sensation for the fastidious listener. If the rms value of the introduced harmonics is kept below 5 per cent of the rms value of the fundamental, the non-linear distortion will not be objectionable (sometimes even 10 per cent is allowed).

One important characteristic of this type of distortion is that, once introduced, it is difficult if not impossible to correct for it, so that the component parts of the circuit must be separately designed so as to have linear voltage versus current characteristics (except for modulators and detectors, see Section 7).

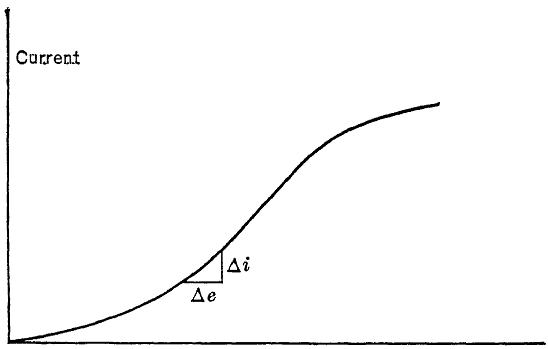


FIG. 1. Non-linear Current-voltage Characteristic

SOLUTION OF NON-LINEAR CIRCUITS. The change in inductance due to the non-linear B-H curve of iron, the varying resistance of an electric arc or of the thermionic vacuum tube, or the varying capacitance of the condenser microphone, all introduce non-linear effects. The simplest case is where the *resistance* is some function of voltage or current. Frequently this functional relationship is given as a curve of current against voltage, as in Fig. 1. The first step is to fit an analytic expression to the curve. Two methods have been widely used in communication practice:

(*a*) a power series of the form

$$i = A_0 + A_1 e + A_2 e^2 + A_3 e^3 + \cdots \tag{2}$$

(*b*) a trigonometric series

$$i = A_0 + A_1 \sin (e + \theta_1) + A_2 \sin (2e + \theta_2) + \cdots \tag{3}$$

16. POWER SERIES SOLUTION

Such devices as thermionic vacuum tubes, electric arcs, some surface contacts (such as copper oxide and lead), and certain crystalline structures have a current-voltage characteristic similar to that shown in Fig. 1. Since the instantaneous value of the electric resistance to an increment of current is $\Delta e/\Delta i$, the resistance varies with the applied voltage, or $r = F(e)$.

The coefficients of the power series of eq. (2) may be evaluated as follows: If four terms of the power series are of interest, choose four separated points on the current-voltage curve and use the four sets of values of e and i in eq. (2). Solve the four resulting algebraic equations by any of the standard methods for A_1, A_2, etc. This method may always be used no matter how irregular the curve and will always give correct results for the points chosen; unless the curve is smooth it may give quite incorrect results for intermediate points. The characteristic must be plotted including all resistance in the circuit whether variable or invariable. It is extremely difficult to include the effect of reactance in the circuit. The application of the method is thus limited, but it is useful in certain special cases.

TAYLOR'S SERIES. If the equation of the current-voltage characteristic is known, or if it can be found (it is frequently of the form $i = Ke^{\eta}$ for at least portions of the curve, in which case the constants can be evaluated by plotting on logarithmic cross-section paper), the solution by Taylor's series is useful in evaluating the coefficients of eq. (2).

The current at any point (e) in terms of the current at a particular point (E_0) can be written.

$$i = F(E_0) + (e - E_0)\left.\frac{\partial i}{\partial e}\right]_0 + \frac{(e - E_0)^2}{2!}\left.\frac{\partial^2 i}{\partial e^2}\right]_0 + \cdots \tag{4}$$

where the derivatives are all evaluated at the operating point (E_0). Care must be exercised in using this formula that the voltage is confined to the region wherein the curve follows the assumed law and within the limits of convergence of the series. A detailed treatment of this method is given in article 20.

17. TRIGONOMETRIC SERIES

In some applications of devices such as vacuum tubes and gas-filled tubes, the devices operate over a range large compared with the part of the characteristic shown in Fig. 1. The application of a power series would usually, under such conditions, require an unreasonable number of terms in order to obtain even a fair approximation. On the other hand, a trigonometric series with a properly chosen fundamental period permits the expansion of such a curve using only a few terms for a fair approximation.

The trigonometric series is periodic by nature and so does not represent the characteristic outside of the interval over which the harmonic analysis was taken. When the curve is of the form abc in Fig. 2 and the voltage varies only over the interval dbe, the curve can be arbitrarily replaced by any other curve outside of this interval. If the curve outside of this interval is replaced by a curve as shown by the broken line in Fig. 2, such that the resulting curve is skew-symmetric about the point e and symmetrical about the vertical axis through d, eq. (3) reduces to

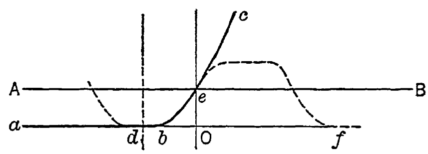


FIG. 2. Construction for Use of Trigonometric Series

$$i = A_0 + A_1 \sin e + A_3 \sin 3e + \cdots \tag{5}$$

Because of the symmetrical properties of the resulting curve both the even harmonics and the phase angles become zero. The choice of curve outside of the operating interval so that the resulting curve has these properties of symmetry is always possible. The convergence of the series is usually greatly increased by using this built-up curve (with a fundamental period four times the distance from d to 0). However, other choices of the curve outside of the working interval are even better in particular cases.

A fundamental period having been chosen and the curve having been completed throughout this period, the coefficients of the harmonic sine functions can be determined by means of one of the several available schemes for harmonic analysis (see article 24).

In most electrical engineering the e has the form of a constant voltage plus a series of sine-wave voltages. When voltages of this form are substituted in eq. (5) and the successive terms expanded by means of the trigonometric formulas for the sine and cosine of the sum of two angles, the series consists of terms of the form cos (x sin d) and sin (x sin d) which can be expanded by the formulas of Jacobi

$$\cos (x \sin d) = J_0(x) + 2\sum_{k=1}^{\infty} J_{2k}(x) \cos 2Rd$$

and

$$\sin (x \sin d) = 2\sum_{R=1}^{\infty} J_{2R-1}(x) \sin (2R - 1)d$$

where $J_n(x)$ is a Bessel function of the first kind, of order n and modulus x. These functions can be found in tables of Bessel functions. See p. 1-37.

By means of this expansion the amplitudes of the sinusoidal components of the current flowing in any circuit can be calculated from the current-voltage characteristic of the circuit, provided that the circuit contains no reactive elements. No general analytic method has yet been devised for the direct application of the trigonometric series in a circuit containing reactive impedance analogous to that treated in article 20. However, eq. (3) or (5) having been obtained, the differential coefficients of eq. (4) can be obtained simply by differentiation.

This solution can be extended to the case of two or more independent variables as in the application to the three-electrode vacuum tube with variable amplification factor and with independent voltages applied in both the grid and plate circuits. However, the work is so voluminous that it is practicable only in very special cases.

18. INDUCTANCE VARIATION

When the variation of inductance is of importance the magnetization curve (locus of the tips of the hysteresis loops) of the iron forming the (closed) magnetic circuit of the reactor is required. This is usually similar in form to the curve in Fig. 1 with flux density (B) as ordinates and magnetizing force (H) as abscissas. This curve must be replotted in the form of i (the exciting current) as a function of ϕ (the total flux). Then by Taylor's formula

$$i = F(\phi_0) + (\phi - \phi_0) \left.\frac{\partial i}{\partial \phi}\right]_0 + \frac{(\phi - \phi_0)^2}{2!} \left.\frac{\partial^2 i}{\partial \phi^2}\right]_0 + \cdots \tag{6}$$

in which the derivatives are evaluated at the operating point. Define

$$\left.\frac{\partial i}{\partial \phi}\right]_0 \equiv \frac{1}{L_0}$$

where L_0 is the usual incremental inductance (apparent inductance to a *change* in current) at the operating point. Then (see article 20)

$$\left.\frac{\partial^2 i}{\partial \phi^2}\right]_0 \equiv - \frac{1}{L_0^2} \left.\frac{\partial L}{\partial \phi}\right]_0$$

and so forth. Also $e = -d\phi/dt$ so that $\int e\, dt = -\phi$. If increments are considered

$$\phi - \phi_0 = \phi_a = -\int e_a\, dt$$

If $e_a = \Sigma E_n \cos(\omega_n t + \theta_n)$ as is usual (see article 21), then

$$\phi_a = -e_a \left[\times \frac{1}{j\omega} \right]$$

where the symbol $[\times]$ indicates that the maximum value of each first-degree cosine term of different frequency in the e immediately preceding is to be multiplied by the modulus of the complex quantity within $[\times]$, and the phase angle of the complex (in this case $\pi/2$) is to be added to the phase of the cosine, both modulus and phase being evaluated at the frequency of the given cosine term (see article 20).

Substituting these expressions in eq. (6)

$$i_a = \frac{-e_a\left[\times \frac{1}{j\omega}\right]}{L_0} - \frac{\left\{e_a\left[\times \frac{1}{j\omega}\right]\right\}^2}{2L_0^2} \left.\frac{\partial L}{\partial \phi}\right]_0 + \cdots \tag{7}$$

Hence if the B-H curve were a straight line $\partial L/\partial \phi = 0$ and

$$i_a = -\sum \frac{E_n \cos(\omega_{nt} + \theta_n - \pi/2)}{\omega_n L_0}$$

To obtain i in terms of the applied voltage E'_a in the circuit containing the variable inductance and a constant impedance z the same process employed in introducing impedance load in the case of a varying resistance must be employed (see article 21)

There results

$$i = e'[\times c_1] + e'^2[\times c_2] + \cdots \tag{8}$$

where

$$c_1 = \frac{1}{z + j\omega L_0} \equiv \frac{1}{z'}$$

$$c_2 = -\frac{L_0}{2} \left.\frac{\partial L}{\partial \phi}\right]_0 \frac{(c_1)^2}{z}$$

which is to be so interpreted that

$$c_{2(m+n)} = -\frac{L_0 c_{1m} c_{1n}}{2z_{(m+n)}} \left.\frac{\partial L}{\partial \phi}\right]_0$$

and

$$c_3 = \frac{c_1(c_2 z)}{z'} \left.\frac{\partial L}{\partial \phi}\right]_0 + \frac{L_0 (c_1)^3}{3! z} \left\{ 2 \left(\left.\frac{\partial L}{\partial \phi}\right]_0 \right)^2 - L_0 \left.\frac{\partial^2 L}{\partial \phi^2}\right]_0 \right\}$$

which is to be so interpreted that

$$\mathrm{c}_3 = \frac{\mathrm{c}_{1m}\mathrm{c}_{2(n+q)}\mathrm{z}_{(n+q)} + \mathrm{c}_{1n}\mathrm{c}_{2(m+q)}\mathrm{z}_{(m+q)} + \mathrm{c}_{1q}\mathrm{c}_{2(m+n)}\mathrm{z}_{(m+n)}}{\mathrm{c}'_{(m+n+q)}} \left.\frac{\partial L}{\partial \phi}\right]_0$$

$$+ \frac{L_0 \mathrm{c}_{1m}\mathrm{c}_{1n}\mathrm{c}_{1q}}{3!\,\mathrm{z}'_{(m+n+q)}} \left\{ 2\left(\left.\frac{\partial L}{\partial \phi}\right]_0\right)^2 - L_0 \left.\frac{\partial^2 L}{\partial \phi^2}\right]_0 \right\}$$

19. CAPACITANCE VARIATION

When the variation of capacitance is of importance, as in such devices as the condenser transmitter, eq. (1) may be rewritten as

$$E = \frac{q}{c} + r\frac{dq}{dt} + L\frac{d^2q}{dt^2} \tag{9}$$

and the characteristic of the device plotted as charge against voltage. The same methods of solution used for resistance variation are then available (with the exception that the operator z introduced in the detailed solution of article 21 must be replaced by $j\omega\mathrm{z}$).

20. APPROXIMATE SERIES EXPANSION FOR THE PLATE CURRENT OF A TRIODE (ASSUMES μ CONSTANT)

Circuits containing thermionic vacuum tubes are chosen to exemplify the detailed treatment of non-linear electric circuits because of the commanding importance of the thermionic tube in communication practice. This wide use has occurred largely because the triode, tetrode, etc., combine an amplification or control function with the non-linearity of the current-voltage device; in most tubes it is possible to employ either or both functions by a simple shift in electrode operating voltages (see Section 4 for nomenclature and characteristics).

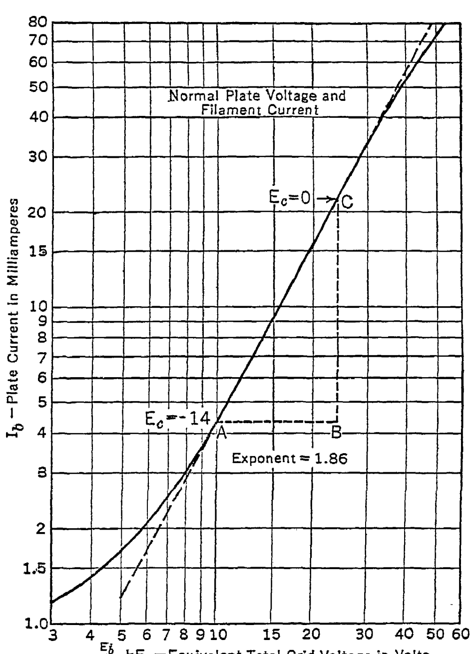


FIG. 3. Mutual Characteristic of a Triode. Normal plate voltage = 130 volts. μ = 5.45.

The triode is generally used in practice with steady voltages applied to both grid and plate, and in addition at least one varying voltage on one of the control electrodes, and frequently with one or more varying voltages impressed in both the plate and grid circuits; these instantaneous currents (i_b, i_c) and voltages (e_b, e_c) may be conveniently split up as follows:

$$i_b = I_b + i_p$$
$$i_c = I_c + i_g$$
$$e_c = E_c + e_g$$
$$e_b = E_b + e_p$$

where the capital letters (E_b, E_c, I_b, I_c) indicate steady values obtaining before the application of any varying voltages in the plate or grid circuits; the lower-case letters with subscript p or g indicate instantaneous values of varying components. All these voltages are specified as actual voltages between electrodes; if the external circuits contain impedance these voltages will differ from the supply voltages.

The plate current of the usual triode is a function of e_c, e_b, and the amplification factor

$\mu(\equiv -\partial e_b \cdot \partial e_c$ when i_b is held constant) such that for at least portions of the operating range

$$i_b = K\left(\frac{e_b}{\mu} + e_c\right)^\eta \tag{10}$$

where K and η are constants. The values may be found for any tube by plotting the plate current of the tube against equivalent grid voltage $\left(\frac{e_b}{\mu} + e_c\right)$ on logarithmic cross-section paper as shown in Fig. 3. The value of η in most triodes usually lies between 1.5 and 2.5; since it varies around 2.0 the triode is sometimes said to follow a square law. This may be sufficiently accurate for rough calculations but is not satisfactory in an exact analysis.

Expanding by Taylor's series

$$i_b = F(E_b, E_c) + (e_p + \mu e_g)\left.\frac{\partial i_b}{\partial e_b}\right]_0 + \frac{(e_p + \mu e_g)^2}{2!}\left.\frac{\partial^2 i_b}{\partial e_b^2}\right]_0 + \frac{(e_p + \mu e_g)^3}{3!}\left.\frac{\partial^3 i_b}{\partial e_b^3}\right]_0 + \cdots \tag{11}$$

where the subscript 0 after a bracket indicates that the derivatives are to be evaluated for $e_b = E_b$ and $e_c = E_c$. Note that $I_b = F(E_b, E_c)$ and may be subtracted from each side of eq. (11). Also define the *static plate resistance* as

$$\frac{1}{r_p} \equiv \left.\frac{\partial i_b}{\partial e_b}\right]_0 \tag{12a}$$

then

$$\left.\frac{\partial^2 i_b}{\partial e_b^2}\right]_0 = -\frac{1}{r_p^2}\left.\frac{\partial r_p}{\partial e_b}\right]_0 \tag{12b}$$

and

$$\left.\frac{\partial^3 i_b}{\partial e_b^3}\right]_0 = \frac{2}{r_p^3}\left(\left.\frac{\partial r_p}{\partial e_b}\right]_0\right)^2 - \frac{1}{r_p^2}\left.\frac{\partial^2 r_p}{\partial e_b^2}\right]_0 \tag{12c}$$

Making these substitutions, eq. (11) becomes

$$i_p = \frac{(e_p + \mu e_g)}{r_p} - \frac{(e_p + \mu e_g)^2}{2r_p^2}\left.\frac{\partial r_p}{\partial e_b}\right]_0 + \frac{2(e_p + \mu e_g)^3}{3!r_p^3}\left(\left.\frac{\partial r_p}{\partial e_b}\right]_0\right)^2$$
$$- \frac{(e_p + \mu e_g)^3}{3!r_p^2}\left.\frac{\partial^2 r_p}{\partial e_b^2}\right]_0 + \cdots \tag{13}$$

If, as is usual in amplifiers, $e_p = 0$, then the factor $\mu/r_p(\equiv \partial i_b/\partial e_c)$ frequently appears in this equation. It is designated g_m and called the *grid-plate transconductance* (mutual conductance). Expansions in terms of transconductance are particularly useful when μ is variable or when the plate resistance of the tube is so high that external resistances are negligible (see article 22).

21. CHARACTERISTICS OF TRIODE WITH LOAD

When the triode is used with external impedance there are voltage drops in these impedances so that the electrode voltages are the differences between the applied voltages and these impedance drops. Since eq. (13) applies to the voltages *between the electrodes,* account must be taken of the impedance drops.

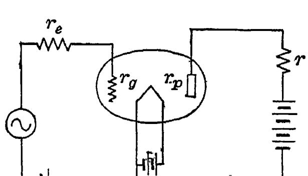


Fig. 4. Triode with Resistance Load

RESISTANCE LOADS. When the external loads are pure resistances the electrode currents and voltages may be split up as follows (see Fig. 4).

$$i_b = I_b + i_p$$
$$i_c = I_c + i_g$$
$$e_c = E_c + e_g = E'_c - r_e I_c + e'_g - r_e i_g$$
$$e_b = E_b + e_p = E'_b - rI_b + e'_p - ri_p$$

where e_p is the sum of the externally impressed voltage and the load resistance drop in the plate circuit, e_g is similar, and r and r_e are the external resistances in the plate and grid circuit. The voltages e'_g and e'_p are the driving voltages introduced in the grid and plate circuits.

If the assumption is made that the plate-grid transconductance $g_n(\equiv \partial i_c/\partial e_b)$ is zero, as is usual for small grid voltages, then the grid current (r_g is the internal grid resistance) is

$$i_g = \frac{e'_g}{r_g + r_e} - \frac{e'^2_g r_g}{2(r_g + r_e)^3}\frac{\partial r_g}{\partial e_c} + \cdots \tag{14}$$

and the varying voltage between grid and cathode is

$$e_g = \frac{e'_g r_g}{r_g + r_e} + \frac{e'^2_g r_e r_g}{2(r_g + r_e)^3}\frac{\partial r_g}{\partial e_c} + \cdots \tag{15}$$

The varying component of the plate current (let $e = e_g + e'_p/\mu$) is

$$i_p = \frac{\mu e}{r_p + r} - \frac{\mu^2 e^2 r_p}{2!(r_p + r)^3}\left.\frac{\partial r_p}{\partial e_b}\right]_0 + \frac{\mu^3 e^3 r_p(2r_p - r)}{3!(r_p + r)^5}\left(\left.\frac{\partial r_p}{\partial e_b}\right]_0\right)^2 - \frac{\mu^3 e^3 r_p^2}{3!(r_p + r)^4}\left.\frac{\partial^2 r_p}{\partial e_b^2}\right]_0 + \cdots \tag{16}$$

Usually the higher-order derivatives rapidly decrease in value, and, since in the higher-order coefficients the power of $(r_p + r)$ in the denominator rises rapidly, all terms beyond the third may be neglected for most practical applications of the tube.

IMPEDANCE LOADS. When impedances are connected in the external circuits of the tube (instead of the resistances of Fig. 4) it is impossible to write the external impedance drop in vector form until the plate circuit is specified. Use may be made of the *operational* impedance $\left(z \equiv r + L\dfrac{d}{dt} + \dfrac{1}{C}\int dt\right)$ which is such that zi *represents* the impedance drop regardless of the form of i. Of course it is impossible to evaluate zi until i is known. The plate current may be written in terms of an operator denoting a delayed multiplication, called the *square cross bracket*, as

$$i_p = e[\times \mathrm{c}_1] + e^2[\times \mathrm{c}_2] + e^3[\times \mathrm{c}_3] + \cdots \tag{17}$$

where

$$\mathrm{c}_1 = \frac{\mu}{r_p + \mathrm{z}} = \frac{\mu}{\mathrm{z}'} \tag{18}$$

in which z is to be separately evaluated for each component frequency of interest. Note that z′ is the total impedance of the plate circuit at the particular frequency. Likewise

$$\mathrm{c}_2 = -\frac{r_p}{2}\left.\frac{\partial r_p}{\partial e_b}\right]_0 \frac{(\mathrm{c}_1)^2}{\mathrm{z}'} \tag{19}$$

in which each of the c's is to be evaluated at the frequency of one of the original frequencies beating together and z′ is to be evaluated at the beat frequency. For instance, to find the size of the current component of periodicity $(\omega_m + \omega_n)$ caused by the beating of two terms of periodicity ω_m and ω_n, evaluate

$$\mathrm{E}_m\mathrm{E}_n\mathrm{c}_{2(m+n)} = -\frac{r_p \mathrm{E}_m \mathrm{E}_n \mu^2}{2!\,\mathrm{z}'_m \mathrm{z}'_n \mathrm{z}'_{(m+n)}}\left.\frac{\partial r_p}{\partial e_b}\right]_0 \tag{19a}$$

Similarly c_3, which embraces terms caused by three of the originally impressed components beating together (these are very important in calculating the amount of interference introduced in a radio receiver by neighboring carriers), is

$$\mathrm{c}_3 = \frac{\mathrm{c}_1(\mathrm{c}_2\mathrm{z})}{\mathrm{z}'}\left.\frac{\partial r_p}{\partial e_b}\right]_0 + \frac{r_p(\mathrm{c}_1)^3}{3!\,\mathrm{z}'}\left\{2\left(\left.\frac{\partial r_p}{\partial e_b}\right]_0\right)^2 - r_p\left.\frac{\partial^2 r_p}{\partial e_b^2}\right]_0\right\} \tag{20}$$

which is to be so interpreted that

$$\mathrm{c}_{3(m+n+q)} = \frac{\mathrm{c}_{1m}\mathrm{c}_{2(n+q)}\mathrm{z}_{(n+q)} + \mathrm{c}_{1n}\mathrm{c}_{2(m+q)}\mathrm{z}_{(m+q)} + \mathrm{c}_{1q}\mathrm{c}_{2(m+n)}\mathrm{z}_{(m+n)}}{\mathrm{z}'_{(m+n+q)}}\left.\frac{\partial r_p}{\partial e_b}\right]_0 + \frac{r_p\mathrm{c}_{1m}\mathrm{c}_{1n}\mathrm{c}_{1q}}{3!\,\mathrm{z}'_{(m+n+q)}}\left\{2\left(\left.\frac{\partial r_p}{\partial e_b}\right]_0\right)^2 - r_p\left.\frac{\partial^2 r_p}{\partial e_b^2}\right]_0\right\} \tag{20a}$$

Table 1. Plate Currents of a Triode

1. No load and resistance load; μ assumed constant.

$$i_p = c_1e + c_2e^2 + c_3e^3 + \text{etc.} \left(e = e_g + \frac{e'_p}{\mu}\right)$$

	c_1	c_2	c_3
No load, eq. (11)	$\frac{\mu}{r_p}$	$-\frac{\mu^2}{2r^2_p}\frac{\partial r_p}{\partial e_b}\Big]_0$	$\frac{\mu^3}{3r^3_p}\left\{\left(\frac{\partial r_p}{\partial e_b}\Big]_0\right)^2 - \frac{r_p}{2}\frac{\partial^2 r_p}{\partial e_b^2}\Big]_0\right\}$
Resistance load, eq. (16)	$\frac{\mu}{r_p + r}$	$-\frac{\mu^2 r_p}{2!(r_p + r)^3}\frac{\partial r_p}{\partial e_b}\Big]_0$	$\frac{\mu^3 r_p}{3!(r_p + r)^5}\left\{(2r_p - r)\left(\frac{\partial r_p}{\partial e_b}\Big]_0\right)^2 - r_p(r_p + r)\frac{\partial^2 r_p}{\partial e_b^2}\Big]_0\right\}$

2. Impedance load; μ assumed constant.

$$i_p = e[\times \mathrm{c}_1] + e^2[\times \mathrm{c}_2] + e^3[\times \mathrm{c}_3] + \text{etc.}$$

c_1	c_2	c_3
$\frac{\mu}{r_p + \mathrm{z}} = \frac{\mu}{\mathrm{z}'}$ eq. (18)	$-\frac{r_p}{2}\frac{\partial r_p}{\partial e_b}\Big]_0 \frac{(\mathrm{c}_1)^2}{\mathrm{z}'}$ eq. (19)	$\frac{\mathrm{c}_1(\mathrm{c}_2\mathrm{z})}{\mathrm{z}'}\frac{\partial r_p}{\partial e_b}\Big]_0 + \frac{r_p}{3}\frac{(\mathrm{c}_1)^3}{\mathrm{z}'}\left\{\left(\frac{\partial r_p}{\partial e_b}\Big]_0\right)^2 - \frac{r_p}{2}\frac{\partial^2 r_p}{\partial e_b^2}\Big]_0\right\}$ eq. (20)

The expression $e^2[\times c_2]$ means that e^2 is to be reduced to *first-degree* cosine terms, after which each cosine term is to be multiplied by the modulus of c_2, evaluated at the frequency or frequencies contained in that term, and the phase angle of c_2 is to be added to the phase of the cosine term. For the method of assigning frequencies in c_2 and c_3 see eqs. (19*a*) and (20*a*).

As in the case of resistance load the grid voltage e_g does not equal the voltage introduced into the grid circuit (e'_g), but can be obtained therefrom, as

$$e_g = e'_g\left[\times \frac{1}{\mathrm{z}''}\right]r_g + e'^2_g\left[\times \frac{\mathrm{z}_e}{(\mathrm{z}'')^2\mathrm{z}''}\right]\frac{r_g}{2}\frac{\partial r_g}{\partial e_c}\Big]_0 + \cdots \tag{21}$$

in which z'' is the total impedance of the grid circuit. In the second term $(\mathrm{z}'')^2$ is to be treated like the term $(\mathrm{c})^2$ in eq. (19), that is evaluated at the separate frequencies which are beating together.

For example, assume a pure cosine voltage $e'_p = E'_{ps}\cos(\omega_s t + \theta_s)$ impressed in the plate circuit, and a similar voltage $e_g = E_{gn}\cos(\omega_n t + \theta_n)$ between grid and filament (or anywhere in the grid circuit if the grid is held negative). Then as far as c_1 and c_2 are concerned

$$\begin{aligned} i_{pa} = {} & \frac{\mu E_{gn}\cos(\omega_n t + \theta'_n)}{\mathrm{z}'_n} + \frac{E'_{ps}\cos(\omega_s t + \theta'_s)}{\mathrm{z}'_s} \\ & - \frac{r_p}{4}\frac{\partial r_p}{\partial e_b}\Big]_0\left[\frac{\mu^2 E_{gn}^2}{z'^2_n(r_p + r)} + \frac{E'^2_{ps}}{z'^2_s(r_p + r)} + \frac{\mu^2 E_{gn}^2\cos(2\omega_n t + \beta_{2n})}{z'^2_n z'_{(2n)}}\right. \\ & + \frac{E'^2_{ps}\cos(2\omega_s t + \beta_{3s})}{z'^2_s z'_{(2s)}} + \frac{2\mu E_{gn}E'_{ps}\cos(\omega_s t + \omega_n t + \beta_{4n})}{z'_s z'_n z'_{(s+n)}} \\ & \left. + \frac{2\mu E_{gn}E'_{ps}\cos(\omega_s t - \omega_n t + \beta_{5n})}{z'_s z'_n z'_{(s-n)}}\right] \end{aligned} \tag{22}$$

where *

$$\theta'_n = \theta_n - \tan^{-1}\frac{x_n}{r_p + r} \text{ and } \theta'_s \text{ is similar}$$

$$\beta_{2n} = 2\theta'_n - \tan^{-1}\frac{x_{(2n)}}{r_p + r} \text{ and } \beta_{3s} \text{ is similar}$$

$$\beta_{4n} = \theta'_n + \theta'_s = \tan^{-1}\frac{x_{(s+n)}}{r_p + r}$$

and

$$\beta_{5n} = \theta'_s - \theta'_n = \tan^{-1}\frac{x_{(s-n)}}{r_p + r}$$

* Note carefully the positive sign associated with $\tan^{-1}\frac{x_n}{r_p + r}$ in β_{5n}. This occurs through the subtraction of θ'_n from θ'_s, even though z_n is in the denominator. The importance of this difference in sign of this term in β_{4n} and β_{5n} is brought out in Section 7, article 18.

A comparison of the expressions for c_1 and c_2 given by eqs. (18) and (19), respectively, shows the effect of the impedance load, which appears in the denominator of c_2 to a higher power, in decreasing the percentage value of the harmonic terms, or in straightening the plate and mutual characteristics. From this it may be seen that an impedance load acts similarly to a resistance load in tending to decrease the relative size of the harmonics introduced by the non-linear shape of these characteristics. The impedance load of course introduces frequency distortion.

DYNAMIC PLATE RESISTANCE. When the steady voltages applied to a triode are such as to cause the static operating point to be within a curved region of the mutual (or other) characteristic, or so near to a curved portion that the applied alternating voltage causes the triode to work on a curved portion for a part of a cycle, more accurate results will be obtained from the series expansion by evaluating the derivatives at the average values of voltages obtaining during the cycle. These may be quite different from the static values. The plate resistance so defined $\left(\frac{1}{r_p} \equiv \frac{\partial i_b}{de_b}\Big]_{\text{average}}\right)$ is called the *dynamic plate resistance*; see eq. (12). All other derivatives should likewise be evaluated at the dynamic operating point, if there is considerable change in the direct, or average, current when the signal is impressed.

The dynamic parameters are quite tedious to compute but their values may be approximately measured on an impedance bridge if the amplitude of the applied voltage is the same as that of the signal voltage. The resistance measured on the bridge is called the *effective plate resistance* and for small or medium voltages is equal to the dynamic plate resistance. When they are not equal the dynamic resistance is the value to be used in evaluating the series; the effective value must be used if it is desired to express all the fundamental current in one term.

VARIATION OF AMPLIFICATION FACTOR. If the variation of μ must be considered (it always should be in pentodes) the first term (c_1), Table 1, is unchanged but the second term (c_2) is

$$c_2 = \frac{-\mu^2 r_p \frac{\partial r_p}{\partial e_b} + r_p(z'_m + z'_n)\frac{\partial \mu}{\partial e_c}}{2z'_m z'_n z'_{(m+n)}}$$

which shows the increased distortion arising from the variation of μ.

22. ANALYSIS FOR MULTI-ELECTRODE TUBES

The above method of analysis can be extended to tetrodes, pentodes, etc. If voltage is introduced into only one control circuit, and if the other grids have low impedance, the analysis as given is sufficient for any case. If, however, there are impedance drops in several of the electrode circuits each must be analyzed in the same manner. The equations are very complicated and not of sufficiently general interest to be included here (see *High-Frequency Alternating Currents*, McIlwain and Brainerd, John Wiley & Sons, Appendix A, for complete development).

23. METHOD OF SUCCESSIVE APPROXIMATIONS *

The calculation of the coefficients in the power series expansion of a vacuum tube and its associated circuit becomes extremely complicated for terms higher than the third term and for analyses of multi-electrode tubes. Thus the analysis of circuits containing vacuum tubes by means of the Taylor's series expansion of the tube characteristics, although theoretically applicable to all circuits, is applicable in practice only to those circuits for which the series converges very rapidly. The factors affecting the convergence are (1) the amplitude of the applied voltages and (2) the sharpness of the curvature of the characteristic in the operating range, that is, the magnitude of the higher derivatives of the characteristic. Thus for high level voltage amplifiers, power amplifiers, large signal detectors (including linear detectors), and most oscillators, the treatment by Taylor's series is practically useless except possibly for qualitative analysis.

In such cases the method of successive approximations is sometimes useful. Equations of the form

$$i_1 = F_1(e_1, e_2) \tag{23a}$$

$$i_2 = F_2(e_1, e_2) \tag{23b}$$

are written for the total number of currents and voltages involved.†

---

* Articles 23 and 24 were contributed by Dr. Carl C. Chambers.

† This method is applicable to certain types of discontinuous functions such as those found in dealing with gas-filled tubes.

For the triode with voltages in both the plate and grid the equations become

$$i_p = F_1(\{e'_g - i_g[\times z_g]\}, \quad \{e'_p - i_p[\times z_p]\}) \tag{24a}$$

$$i_g = F_2(\{e'_g - i_g[\times z_g]\}, \quad \{e'_p - i_p[\times z_p]\}) \tag{24b}$$

where the symbols have the same meaning as in the previous section on the Taylor's series expansion of vacuum tubes.

The values of the resulting currents are estimated and called i_{p0} and i_{g0}. These can be obtained from oscillograms of a similar vacuum tube, from the approximate solution by means of the Taylor's series expansion, or simply from an intuitive guess. Although the accuracy of the estimate of the value of i_{p0} and i_{g0} is not theoretically important, the labor involved is directly dependent upon this accuracy.

Having i_{p0} and i_{g0}, the first approximation is calculated by substitution of these values in the right-hand members of (24*a*) and (24*b*), giving

$$i_{p1} = F_1(\{e'_g - i_{g0}[\times z_g]\}, \quad \{e'_p - i_{p0}[\times z_p]\})$$

$$i_{g1} = F_1(\{e'_g - i_{g0}[\times z_g]\}, \quad \{e'_p - i_{p0}[\times z_p]\})$$

The second and higher order approximations are obtained by substituting the preceding approximation in the right-hand members of eqs. (24).

In making these successive substitutions it is necessary to write the preceding approximation in separate sine-wave terms in order to evaluate $i_{gr}[\times z_g]$ and $i_{pr}[\times z_p]$. This can be done by any of the various methods for harmonic analysis (see Fisher-Hinnen, *Method and Wave Analysis*, article 1). When z_p and z_g are selective impedances, the only harmonic components of the current of importance for purposes of substitution are those for which z_p and z_g are not essentially zero. For this reason this method of analysis is especially applicable to class B and class C r-f amplifiers, oscillators, and frequency multipliers, where the plate impedances are highly selective.

The functions F_1 and F_2 can be in any form in which they are completely specified over the entire operating range of the applied voltages. Thus any analytical expression for the tube currents or any complete set of curves is sufficient for use in the above method of analysis.

24. HARMONIC ANALYSIS OF THE CURRENT FOR A SINUSOIDAL APPLIED VOLTAGE

In many cases the performance of a non-linear circuit can be predicted from a knowledge of the harmonic content of the current when a sinusoidal voltage is impressed in the circuit. The usual discussion of the merit of a non-selective amplifier assumes a sine-wave excitation, to avoid the complexity introduced by the presence of the many beat terms otherwise present.

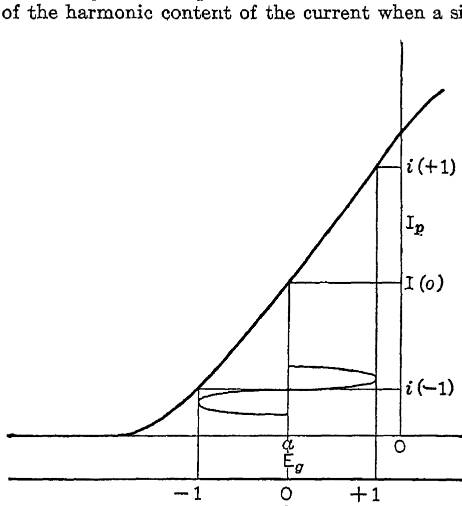


FIG. 5. Tube Characteristic and Sine-wave Input

When the relation between the instantaneous impressed voltage and the instantaneous current is given it is of course possible to plot the wave form of the current. This may then be analyzed by any of the usual methods of wave analysis.

When the input voltage varies sinusoidally, the output current will be periodic, having a fundamental periodicity equal to that of the input voltage. In resistive circuits the relation between the input voltage and the output current is independent of time, so that the output current will pass from maximum to minimum and from minimum to maximum over the same path, in fact, over the characteristic curve. (The term characteristic is applied in its original sense to the instantaneous relation between input voltage and output current). It follows that, if a complete period of the plate current is T seconds, $i(t)$

$= i(T - t)$ when the origin of time is chosen at the instant when the current is a maximum. This form of symmetry insures that when the output current due to a sinusoidal input voltage is written in the form

$$i = I_0 + I_1 \cos(\omega t + \beta_1) + I_2 \cos(2\omega t + \beta_2) + \qquad (25)$$

the phase angles β_1, β_2, etc., will be zero.

Several methods of analysis are used to obtain the values I_0, I_1, etc. The one of broadest application is the variation of the Fisher-Hinnen method applicable to such a characteristic curve instead of to the wave itself. The analysis by this method can be stated as follows: consider the zero point of the sinusoidal input voltage variation to be the zero of the base line of the characteristic, and consider the scale along this base line to be such that the peak value of the input voltage is unity. Then $i(e)$ is the relation given by the mutual characteristic where e is the input voltage which on this scale varies from -1 to $+1$ and i is the corresponding output current (see Fig. 5). Then the I's to the second approximation in the Fisher-Hinnen method of analysis are given by

$$I_1 = {}^1/_2\,[i(1) - i(-1)]$$
$$I_2 = {}^1/_4\,[i(1) - 2i(0) + i(-1)]$$
$$I_3 = {}^1/_6\,[i(1) - 2i(0.5) + 2i(-0.5) - i(-1)]$$
$$I_4 = {}^1/_8\,[i(1) - 2i(0.707) + 2i(0) - 2i(-0.707) + i(-1)]$$
$$\cdots\cdots\cdots\cdots$$
$$I_n = \frac{1}{2n}\left[i(1) - 2i\left(\cos\frac{\pi}{n}\right) + 2i\left(\cos\frac{2\pi}{n}\right) - 2i\left(\cos\frac{3\pi}{n}\right)\cdots \pm 2i\left(\cos\frac{(n-1)\pi}{n}\right) \mp i(-1)\right]$$

The I's to the fourth approximation in the Fisher-Hinnen method are then given by

$$I'_n = I_n - I_{2n} - I_{3n}$$

The d-c component I_0 is given to the second approximation by

$$I_0 = {}^1/_3\left[i(0) + \frac{i(1) + i(-1)}{2} + \frac{i(0.707) + i(-0.707)}{2}\right]$$

Another method (see D. C. Espley, *Proc. I.R.E.*, Vol. 21, 1439 [1933]) has the advantage that the points along the input voltage axis of the mutual characteristic, at which the current is evaluated, are equally spaced. This analysis using three points gives results identically the same as the values for I_1 and I_2 given by the Fisher-Hinnen method. When five points are used the I's become

$$I_0 = {}^1/_6\,[i(1) + 2i(0.5) + 2i(-0.5) + i(-1)]$$
$$I_1 = {}^1/_3\,[i(1) + i(0.5) - i(-0.5) - i(-1)]$$
$$I_2 = {}^1/_4\,[i(1) - 2i(0) + i(-1)]$$
$$I_3 = {}^1/_6\,[i(1) - 2i(0.5) + 2i(-0.5) - i(-1)]$$
$$I_4 = {}^1/_{12}\,[i(1) - 4i(0.5) + 6i(0) - 4i(-0.5) + i(-1)]$$

For the values of the I's using seven equally spaced points see the original paper referred to above.

In the special case when the even harmonics are balanced out as in "back to back" amplifiers (pushpull class A, class B audio, and class AB amplifiers, see Amplifiers, Section 7) the mutual characteristic becomes like that shown in Fig. 6. I_0 is balanced out of the output current as well as the even harmonics so that $i(e) = -i(-e)$. For such amplifiers the analysis by the Fisher-Hinnen method gives for the remaining I's

$$I_1 = i(1)$$
$$I_3 = {}^1/_3\,[i(1) - 2i(0.5)]$$
$$I_5 = {}^1/_5\,[i(1) - 2i(0.809) + 2i(0.309)]$$
$$I_7 = {}^1/_7\,[i(1) - 2i(0.901) + 2i(0.624) - 2i(0.222)]$$
$$\cdots\cdots\cdots\cdots$$
$$I_n = \frac{1}{n}\left[i(1) - 2i\left(\cos\frac{\pi}{n}\right) + 2i\left(\cos\frac{2\pi}{n}\right) - \cdots \pm 2i\left(\cos\frac{(n-1)\pi}{2n}\right)\right]$$

As before, a better approximation for I_n is obtained by subtracting I_{3n} from the calculated I_n.

When this symmetry exists the method of Espley described above reduces to the Fisher-Hinnen method when the measurements are made at only five points (for this special case only two points need actually be measured), and the seven measured point analysis of Espley reduces to

$$I_1 = {}^1/_{320}\,[167i(1) + 252i(0.667) - 45i(0.333)]$$
$$I_3 = {}^1/_{128}\,[45i(1) - 36i(0.667) - 63i(0.333)]$$
$$I_5 = {}^{81}/_{640}\,[i(1) - 4i(0.667) + 5i(0.333)]$$

Here because of the symmetrical character of the curve only three points actually need be measured.

Mouromtseff and Kozanowski (*Proc. I.R.E.*, Vol. 22, 1090 [1934]) give a somewhat simpler method to calculate up to the eleventh harmonic in the case of a symmetrical curve such as Fig. 6. The straight line *aob* is drawn intersecting the curve at the point of maximum input voltage, that is, $e = 1$ in the notation used above. Then, instead of measuring $i(e)$, $\Delta i(e)$ is measured where $\Delta i(e)$ is the difference in current between the curve and the line *aob* and is taken positive when the line *aob* is above the curve. This gives for the I's in the order in which they are to be calculated

$$I_5 = 0.4[\Delta i(0.309) - \Delta i(0.809)]$$
$$I_3 = 0.4475[\Delta i(0.309) + \Delta i(0.809)] + 0.333\Delta i(0.5) - 0.578\Delta i(0.866) - 0.5I_5$$
$$I_7 = 0.4475[\Delta i(0.309) + \Delta i(0.809)] - I_3 + 0.5I_5$$
$$I_9 = I_3 - 0.667\Delta i(0.5)$$
$$I_{11} = 0.707\Delta i(0.707) - I_3 + I_5$$
$$I_1 = i(1) - I_3 + I_5 - I_7 + I_9 - I_{11}$$

Any of the above methods can be used for the calculation of the I's in eq. (25) when the input voltage is sinusoidal. The per cent amplitude for any given harmonic is then I_n/I_1 for an input voltage, the peak value of which is taken $e = 1$. The complete equation of the curve from $e = -1$ to $e = +1$ can be written in the form

$$i(e) = I_0 + I_1 e + I_2 \cos 2(\cos^{-1} e) + I_3 \cos 3(\cos^{-1} e) + \cdots$$

This expression for the output current is unique; that is, for each value of e between -1 and $+1$ this equation gives the corresponding instantaneous plate current provided that the conditions prescribed at the beginning of this section are fulfilled, chiefly, that the load is essentially a constant resistance over the operating frequencies and voltages. This expression can be used in several ways although calculations of the operation for input voltages other than a simple sinusoid of amplitude unity on the scale of e are extremely complex analytically.

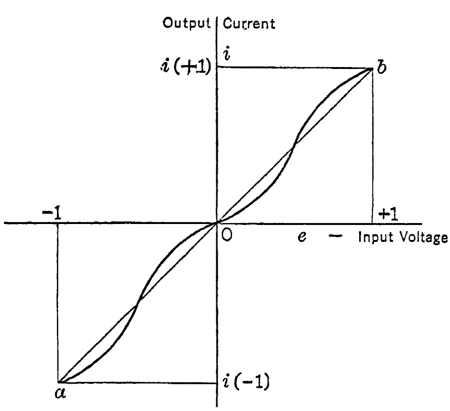


FIG. 6. Mutual Characteristics of Two Tubes "Back to Back"

The calculation of the distortion and output by any of the above methods for a sinusoidal input voltage having an amplitude of a fixed fraction of e can be made by means of this expression for $i(e)$ without again using the curve. Formulas can be developed for this purpose using any of the above methods of analysis as a basis giving the I's corresponding to an input voltage of fractional amplitude.

This method of analysis can generally be used when the resistive load is coupled into the non-linear circuit by a transformer. For if the transformer is "perfect" the relation between voltage and current is the same as for a resistance load (see Transformers, Section 6); in other words the transformer offers a resistive impedance to the circuit. Practical transformers are nearly perfect enough to assume that their only effect is to multiply (or divide) the load resistance by the square of the turn ratio.

25. INPUT IMPEDANCE OF A TRIODE

Neglecting the leakage resistances between electrodes the equivalent circuit of a triode with grid negative is shown in Fig. 7. The equivalent input resistance is

$$r_i = \frac{r_pC_3\left[(r^2+x^2)(\mu C_2 + C_3 + \mu C_3) + r_p r C_3 - \dfrac{\mu x}{\omega}\right]}{(\omega^2 r_p^2 A^2 + B^2)(r^2+x^2) + r_p(C_1+C_3)[2rB - 2\omega r_p x A + r_p(C_1+C_3)]} \tag{26a}$$

the equivalent input reactance by

$$x_i = -\frac{\left\{\begin{array}{l}\omega^2 r_p^2 (C_2+C_3)(r^2+x^2)A + B(r^2+x^2+rr_p)\\ +(C_1+C_3)(r_p^2+rr_p) - \omega r_p^2 x(2A + C_3^2)\end{array}\right\}}{\left\{\begin{array}{l}\omega\{(\omega^2 r_p^2 A^2 + B^2)(r^2+x^2)\\ +r_p(C_1+C_3)[2rB - 2\omega r_p x A + r_p(C_1+C_3)]\}\end{array}\right\}} \tag{26b}$$

where $A = C_1C_2 + C_1C_3 + C_2C_3$, and $B = C_1 + C_2 + \mu C_3$.

Both the numerator and denominator of each of these expressions contain negative terms; thus either can be positive or negative depending on the circuit parameters. If the value of the reactance becomes positive it indicates that the equivalent input reactance is inductive. If the value of the expression for resistance becomes negative, it indicates that the real part of the input impedance of the tube is an *equivalent negative resistance* and is supplying instead of absorbing power. The input resistance will become negative only for certain values of *inductive load;* cf. eq. (26*a*). When this condition exists, the tube will always supply a greater plate current than it would for the same grid voltage and a non-reactive load. When this occurs the tube is said to be *regenerating* (see Section 8). If the negative input resistance exactly equals the external resistance the total resistance of the circuit is zero and current can flow without a driving voltage; in such a case the tube is said to *oscillate* (see Section 7, oscillators).

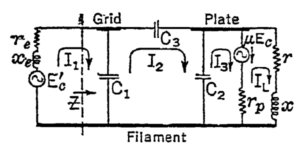


FIG. 7. Simplified Equivalent Circuit of a Triode

When the grid voltage is positive, grid current flows and the grid resistance is defined by $1/r_g = \partial i_g/\partial e_g$. This is practically infinite when the grid is negative but may be quite small when the grid is positive; in this latter case the interelectrode impedances are usually negligible in comparison to it so that the input impedance of the triode is simply r_g.

BIBLIOGRAPHY

Byerly, W. E., *Elements of Integral Calculus*, p. 332. G. E. Stechert and Co.
Carson, J. R., A Theoretical Study of the Three-Element Vacuum Tube, *Proc. I.R.E.*, April 1919.
Steinmetz, C. P., *Engineering Mathematics*, Chapter VI. McGraw-Hill.
McIlwain, K., and J. G. Brainerd, *High-frequency Alternating Currents*, Chapter VI. John Wiley.
Terman, F. E., *Radio Engineering*. McGraw-Hill (1932).
Llewellyn, F. B., Operation of Thermionic Vacuum Tube Circuits, *B.S.T.J.*, Vol. V, 433 (1926).
Barrow, W. L., Contribution to the Theory of Non-linear Circuits with Large Applied Voltages, *Proc. I.R.E.*, Vol. 22, No. 8, 964 (August 1934).
Byerly, W. E., *Elementary Treatise on Fourier Series and Spherical, Cylindrical, and Ellipsoidal Harmonics*. Ginn (1902).
Jahnke, E., and F. Emde, *Tables of Functions with Formulae and Curves*, 2nd Ed., p. 242. B. G. Teubner, Berlin (1933).

ELECTROMAGNETIC RADIATION

By Knox McIlwain

All forms of wireless or radio transmission depend on the fact that electromagnetic energy is radiated from any wire in which a varying current flows. The current in the wire sets up magnetic and electric fields, in which it is usually assumed that the energy associated with the current flow is stored; a portion of this energy, however, is not stored, but continually travels away from the wire, or is radiated. Radiation is a phenomenon which is totally negligible at low frequencies.

26. MAXWELL'S EQUATIONS

There are two well-known experimental relations between magnetic and electric fields. The two fields are interdependent, a change in one always being accompanied by a change of the other. The two laws are:

1. Whenever the net magnetic flux linking any closed loop changes, an electromotive force is set up in the loop. If ϕ is the net magnetic flux linking the loop, then (Faraday's law) the electromotive force is

$$e = -\frac{d\phi}{dt} \tag{1}$$

2. The resultant magnetomotive force (in ampere-turns) acting around any closed loop is equal to the rate of flow of electricity through the surface bounded by the loop plus the time rate of change of electric flux through this surface. If $i = dq/dt$ is the rate of flow of electricity (total conduction current in amperes) and ψ the electric flux through the surface then the magnetomotive force is

$$m = i + \frac{d\psi}{dt} \tag{2}$$

Both these relations are such that the force expressed by the left side of the equation bears a right-hand-screw relation to a positive increment of the right-hand side. They apply to any medium whatever, to conductors, dielectrics, ferromagnetic substances, etc.

By applying these equations to an elementary volume in space, Maxwell's laws of electromagnetism are developed. Stated in vector notation these are (all quantities are expressed in mks units)

$$\gamma \boldsymbol{\varepsilon} + K\frac{\partial \boldsymbol{\varepsilon}}{\partial t} = \text{curl } \boldsymbol{H} \tag{3}$$

$$-\mu\frac{\partial \boldsymbol{H}}{\partial t} = \text{curl } \boldsymbol{\varepsilon} \tag{4}$$

From these the continuity equations immediately follow;

$$\text{div } \boldsymbol{u} = 0 \tag{5}$$

$$\text{div } \boldsymbol{B} = 0 \tag{6}$$

where $\boldsymbol{u}$ is the *total* current density in amperes per square meter and $\boldsymbol{B}$ the magnetic flux density in webers per square meter.

Stated in the ordinary component form these equations are:

$$\gamma \varepsilon_x + \frac{K\partial \varepsilon_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} \tag{3a}$$

$$\gamma \varepsilon_y + \frac{K\partial \varepsilon_y}{\partial t} = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} \tag{3b}$$

$$\gamma \varepsilon_z + \frac{K\partial \varepsilon_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \tag{3c}$$

$$-\mu\frac{\partial H_x}{\partial t} = \frac{\partial \varepsilon_z}{\partial y} - \frac{\partial \varepsilon_y}{\partial z} \tag{4a}$$

$$-\mu\frac{\partial H_y}{\partial t} = \frac{\gamma \varepsilon_x}{\partial z} - \frac{\partial \varepsilon_z}{dx} \tag{4b}$$

$$-\mu\frac{\partial H_z}{\partial t} = \frac{\partial \varepsilon_y}{\partial x} - \frac{\partial \varepsilon_x}{\partial y} \tag{4c}$$

$$\frac{\partial}{\partial_x}\left(\sigma_x + \frac{\partial D_x}{\partial t}\right) + \frac{\partial}{\partial y}\left(\sigma_y + \frac{\partial D_y}{\partial t}\right) + \frac{\partial}{\partial z}\left(\sigma_z + \frac{\partial D_z}{\partial t}\right) = 0 \tag{5a}$$

where σ is the conduction current density in amperes per square meter and $\boldsymbol{D}$ the electric flux density in coulombs per square meter, and

$$\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \tag{6a}$$

Only fields which satisfy these relations are possible, and these equations may be used to determine the necessary form for particular fields.

The energy flow associated with the electromagnetic wave is found by developing Poynting's vector

$$P = \mathcal{E}H \sin(\boldsymbol{\mathcal{E}}, \boldsymbol{H}) \tag{7}$$

which is a vector perpendicular to $\mathcal{E}$ and H, whose magnitude depends on the product of $\mathcal{E}$ and H and the sine of the angle between them (measured counterclockwise from $\boldsymbol{\mathcal{E}}$ to $\boldsymbol{H}$). The integral of Poynting's vector over a surface gives the rate at which energy flows through the surface.

WAVE EQUATION. In an isotropic insulating medium no conduction current can flow, since γ (eq. 3) is zero. Elimination of $\boldsymbol{H}$ between eqs. (3 and 4), and similarly for $\boldsymbol{\mathcal{E}}$, gives the wave equations

$$\frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial t^2} = \frac{1}{\mu K}\left(\frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial x^2} + \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial y^2} + \frac{\partial^2 \boldsymbol{\mathcal{E}}}{\partial z^2}\right) \equiv \frac{1}{\mu K}\nabla^2 \boldsymbol{\mathcal{E}} \tag{8}$$

$$\frac{\partial^2 \boldsymbol{H}}{\partial t^2} = \frac{1}{\mu K}\left(\frac{\partial^2 \boldsymbol{H}}{\partial x^2} + \frac{\partial^2 \boldsymbol{H}}{\partial y^2} + \frac{\partial^2 \boldsymbol{H}}{\partial z^2}\right) \equiv \frac{1}{\mu K}\nabla^2 \boldsymbol{H} \tag{9}$$

Solution of these equations shows that each component of electric or magnetic field must have the form

$$\mathcal{E}_x = f_1\left(t - \frac{x}{v}\right) + f_2\left(t + \frac{x}{v}\right) \tag{10}$$

where $v = \pm 1/\sqrt{\mu K}$. These functions represent incident and reflected waves (see p. 5-23) traveling at the velocity v ($= 3 \times 10^8$ meters per sec in vacuum and close to that in air). The *wavelength* is defined as the distance traveled in one period.

27. PROGRESSIVE PLANE WAVES

The simplest form of wave mathematically is one which travels along the x axis, say, and in which it is assumed that none of the fields vary with y or z. In such case (by Maxwell's laws)

$$\mathcal{E}_x = 0 \tag{11a}$$

$$\mathcal{E}_y = \bar{\mathcal{E}}_y \cos\left[\omega\left(t - \frac{x}{v}\right)\right] \tag{11b}$$

$$\mathcal{E}_z = \bar{\mathcal{E}}_z \cos\left\{\left[\omega\left(t - \frac{x}{v}\right)\right] + \delta_z\right\} \tag{11c}$$

$$H_x = 0 \tag{11d}$$

$$H_y = -\sqrt{\frac{K}{\mu}}\,\mathcal{E}_z = -\sqrt{\frac{K}{\mu}}\,\bar{\mathcal{E}}_z \cos\left\{\left[\omega\left(t - \frac{x}{v}\right)\right] + \delta_z\right\} \tag{11e}$$

$$H_z = \sqrt{\frac{K}{\mu}}\,\mathcal{E}_y = \sqrt{\frac{K}{\mu}}\,\bar{\mathcal{E}}_y \cos\left[\omega\left(t - \frac{x}{v}\right)\right] \tag{11f}$$

in which $\omega = 2\pi f$. Then $\omega[t - (x/v)] = \omega t - 2\pi x/\lambda$.

In the general form the loci of the vectors $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{H}$ are ellipses, so that the plane wave above is said to be *elliptically polarized.* If $\bar{\mathcal{E}}_y = \bar{\mathcal{E}}_z$ and $\delta_z = \pi/2$ or $3\pi/2$ the ellipse reduces to a circle and the wave is said to be *circularly polarized.*

If either $\bar{\mathcal{E}}_y$ or $\bar{\mathcal{E}}_z$ is zero, or if $\delta_z = 0$ or π, the ellipse reduces to a straight line, in which case the wave is called a plane-polarized wave. The electric energy per meter^3 of such a wave at any point in the field is $K\mathcal{E}^2/2$, and the magnetic energy per meter^3 is $\mu H^2/2$. It follows that for a progressive plane-polarized wave at every point the total energy is always half electric and half magnetic, but its value varies, of course, with position and time. The total energy (W) of the field is

$$W = We + W_m = K\int\int\int \mathcal{E}^2\,dr \tag{12}$$

where dr is an element of volume. The average energy is $W_1 = K\mathcal{E}^2/2$, and the average *rate* at which energy passes through a meter^2 in the YZ plane is

$$P_a = \frac{1}{2}\sqrt{\frac{K}{\mu}}\,\bar{\mathcal{E}}_y{}^2 \tag{13}$$

28. FIELDS DUE TO A CURRENT IN A WIRE

If a current i flows in a short length of wire (see Fig. 1) located in free space and assumed at the origin of coordinates and coincident with the z axis, the fields at any point $P(x, y, z)$ are

$$\varepsilon_x = \frac{1}{K}\frac{\partial^2 w}{\partial x\, \partial z} \tag{14a}$$

$$\varepsilon_y = \frac{1}{K}\frac{\partial^2 w}{\partial y\, \partial z} \tag{14b}$$

$$\varepsilon_z = -\frac{1}{K}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right) \tag{14c}$$

$$H_x = \frac{\partial^2 w}{\partial y\, \partial t} \tag{14d}$$

$$H_y = -\frac{\partial^2 w}{\partial x\, \partial t} \tag{14e}$$

$$H_z = 0 \tag{14f}$$

where $w \equiv \frac{1}{d} f_1\left(t - \frac{d}{v}\right)$; in this expression d is the distance from the origin to the point, and $F_1(t)$ expresses the distribution of charge in the wire, so that $F(t) = i\delta l$, where δl is the length of the wire (assuming uniform current for the differential length δl).

If the current is assumed sinusoidal, so that $i = \sqrt{2}I \sin(\omega t + 90°)$ amperes, at any point distant from the wire (so that $d \gg \lambda/2\pi$), the fields are

$$\varepsilon = \frac{\pi i\, \delta l}{150\, d\lambda} \cos\theta \cos(\omega t - \beta d) \text{ volts per meter} \tag{15a}$$

$$H = \frac{0.08\pi^2 i\, \delta l}{d\lambda} \cos\theta \cos(\omega t - \beta d) \text{ ampere-turns per meter} \tag{15b}$$

where θ is the angle of elevation of the point P from a plane perpendicular to the wire. All lengths are expressed in meters.

The direction of the electric field is perpendicular to d and in the plane formed by d and the axis of the wire. The magnetic field is perpendicular to d and the electric field.

Such an elementary wire is called an electric doublet. The fields due to a long length of wire l can be obtained by integration of the elementary length (see Antennas, Section 6.)

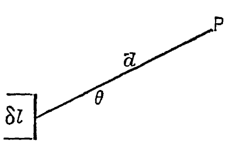


FIG. 1. Elementary Radiator

The average rate of flow of energy per square meter through an area perpendicular to d is (in watts)

$$P_a = \frac{30\pi l^2 I^2}{\lambda^2 r^2} \cos^2\theta \tag{16}$$

and the total power radiated (in watts) is

$$P_T = \frac{80\pi^2 l^2 I^2}{\lambda^2} \tag{17}$$

The radiated power for a given current varies directly as $1/\lambda^2$ or as the square of the frequency. For a given total charge, or a given voltage difference between the ends of the antenna, the radiated power varies as $1/\lambda^4$ or as the fourth power of the frequency.

29. REFLECTION AND REFRACTION

Whenever an electromagnetic wave meets a boundary between two media of different dielectric or magnetic properties, there is a change in the fields and frequently the wave splits into two waves, one of which is *reflected* back into the first medium, the other is *refracted* into the second medium.

CONDITIONS AT THE BOUNDARY. The following proposition is assumed: In two media, the components of the electric (or magnetic) intensity *tangent to the surface separating those media* are equal (in magnitude and direction) at this surface. It follows from this that the normal component (u_N) of the *total* current density is the same on both sides of the surface, that is, $u_N = u_{2N}$, or for insulating media

$$D_N = D_{2N} \tag{18}$$

Similarly

$$B_N = B_{2N} \tag{19}$$

The general expression for a plane wave as far as $\mathbf{\varepsilon}$ is concerned is:

$$\varepsilon_x = \bar{\varepsilon}_x \cos \omega \left(t - \frac{p}{v}\right) \tag{20a}$$

$$\varepsilon_y = \bar{\varepsilon}_y \cos \left[\omega \left(t - \frac{p}{v}\right) + \delta_y\right] \tag{20b}$$

$$\varepsilon_z = \bar{\varepsilon}_z \cos \left[\omega \left(t - \frac{p}{v}\right) + \delta_z\right] \tag{20c}$$

where δ_y and δ_z are constants and $\bar{\varepsilon}_x$ is the maximum value of the X component of $\mathbf{\varepsilon}$.

TWO ISOTROPIC DIELECTRICS. Consider two media separated by a plane surface, and assume that a plane wave in medium 1 impinges on this surface (see Fig. 2). Assume that the total electric intensity $\mathbf{\varepsilon}$ in the first medium is composed of two parts $\mathbf{\varepsilon}_1$ and $\mathbf{\varepsilon}'_1$; then

$$\mathbf{\varepsilon} = \mathbf{\varepsilon}_1 + \mathbf{\varepsilon}'_1 \tag{21}$$

$\mathbf{\varepsilon}_1$ will be called the *incident* wave (see also p. 5-23) and $\mathbf{\varepsilon}'_1$ the *reflected* wave; $\mathbf{\varepsilon}'_1$ will have components similar to those of eq. (20) denoted by primes. Note that, p'_1 being undetermined, what is called the reflected wave is not restricted to travel in a direction directly opposite the incident wave, as is the case on transmission lines.

Assume that the resultant electric intensity in the second medium is $\mathbf{\varepsilon}_2$, called the *refracted* wave, with similar components identified by the subscript 2.

Applying Maxwell's equations at the boundary, $\beta_1 v_1 = \beta'_1 v'_1 = \beta_2 v_2$ or $\omega_1 = \omega'_1 = \omega_2$, so that all waves have the same frequency.

Also the incident, reflected, and refracted waves all have wave normals (perpendiculars to the wave front) in the same plane. Similarly if ϕ is the angle between the wave normal and the normal to the boundary then $\phi_1 = \phi'_1$, or the angle of incidence is equal to the angle of reflection. Also

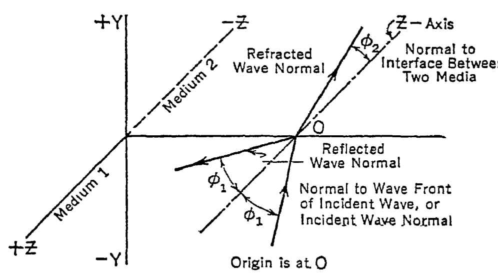


FIG. 2. Wave Normals

$$\frac{\sin \phi_1}{\sin \phi_2} = \frac{v_1}{v_2} = \eta = \frac{\sqrt{\mu_2 K_2}}{\sqrt{\mu_1 K_1}} \tag{22}$$

where η is the relative *index of refraction* of medium 2 with respect to medium 1.

The magnetic intensities of the various waves are, in magnitude (from Maxwell's equations), $H_1 = \sqrt{K_1/\mu_1}\,\varepsilon_1$, $H'_1 = \sqrt{K_1/\mu_1}\,\varepsilon'_1$, and $H_2 = \sqrt{K_2/\mu_2}\,\varepsilon_2$. All these are vectors in the wave fronts such that $\mathbf{\varepsilon}_1$, H_1, and p_1 are mutually perpendicular in such orientation that rotation from $\mathbf{\varepsilon}_1$ to H_1 (through the smaller angle) would move a right-handed screw in the direction of p.

Normal Components. It is convenient to split the electric and magnetic intensities into *two* components, one normal to the *plane of incidence* (not the interface between the media) and one in the plane of incidence. Maxwell's equations apply to each separately, since the equations are assumed linear (there has been some indication, the intermodulation of waves in space, that the equations are not entirely linear). The normal components, at the interface in terms of the incident wave, are

$$\frac{\varepsilon'_{1N}}{\varepsilon_{1N}} = \frac{\bar{\varepsilon}'_{1N}}{\bar{\varepsilon}_{1N}} \underline{/\delta'_{1N}} = \frac{\cos \phi_1 \sqrt{\mu_2 K_1} - \cos \phi_2 \sqrt{\mu_1 K_2}}{\cos \phi_1 \sqrt{\mu_2 K_1} + \cos \phi_2 \sqrt{\mu_1 K_2}} \tag{23a}$$

$$\frac{\varepsilon_{2N}}{\varepsilon_{1N}} = \frac{\bar{\varepsilon}_{2N}}{\bar{\varepsilon}_{1N}} \underline{/\delta_{2N}} = \frac{2 \cos \phi_1 \sqrt{\mu_2 K_1}}{\cos \phi_1 \sqrt{\mu_2 K_1} + \cos \phi_2 \sqrt{\mu_1 K_2}} \tag{23b}$$

Since $\cos\phi_2 = \sqrt{1 - \sin^2\phi_2}$ and since $\sin\phi_1/\sin\phi_2 = \eta$ (the index of refraction), then $\cos\phi_2 = \sqrt{1 - \sin^2\phi_1/\eta}$. Hence, if $\sin\phi_1 > \eta$, $\cos\phi_2$ will be imaginary. This is the case of *total reflection* (see below). The angle of incidence for which $\sin\phi_1 = \eta$ is called the *critical angle.*

Parallel Components. The parallel components at the interface, in terms of the incident wave, are

$$\frac{\mathcal{E}'_{1P}}{\mathcal{E}_{1P}} = \frac{\bar{\mathcal{E}}'_{1P}}{\bar{\mathcal{E}}_{1P}} \underline{/\delta'_{1P}} = \frac{\cos\phi_1\sqrt{\mu_1 K_2} - \cos\phi_2\sqrt{\mu_2 K_1}}{\cos\phi_1\sqrt{\mu_1 K_2} + \cos\phi_2\sqrt{\mu_2 K_1}} \tag{24a}$$

$$\frac{\mathcal{E}_{2P}}{\mathcal{E}_{1P}} = \frac{\bar{\mathcal{E}}_{2P}}{\bar{\mathcal{E}}_{1P}} \underline{/\delta_{2P}} = \frac{2\cos\phi_1\sqrt{\mu_2 K_1}}{\cos\phi_1\sqrt{\mu_1 K_2} + \cos\phi_2\sqrt{\mu_2 K_1}} \tag{24b}$$

Case of Total Reflection. If $\cos\phi_2$ is imaginary let $\cos\phi_2 = jn_2$, where n_2 is the absolute value of the imaginary. Substituting this in eqs. (23) and (24), the normal components are

$$\frac{\mathcal{E}'_{1N}}{\mathcal{E}_{1N}} = \frac{\bar{\mathcal{E}}'_{1N}}{\bar{\mathcal{E}}_{1N}} \underline{/\delta'_{1N}} = \frac{\cos\phi_1\sqrt{\mu_2 K_1} - jn_2\sqrt{\mu_1 K_2}}{\cos\phi_1\sqrt{\mu_2 K_1} + jn_2\sqrt{\mu_1 K_2}}$$

$$= 1 \underline{/2\tan^{-1} - \frac{-n_2\sqrt{\mu_1 K_2}}{\cos\phi_1\sqrt{\mu_2 K_1}}} \tag{25}$$

and the parallel components are

$$\frac{\mathcal{E}'_{1P}}{\mathcal{E}_{1P}} = \frac{\bar{\mathcal{E}}'_{1P}}{\bar{\mathcal{E}}_{1P}}\, \delta'_{1P} = \frac{\cos\phi_1\sqrt{\mu_1 K_2} - jn_2\sqrt{\mu_2 K_1}}{\cos\phi_1\sqrt{\mu_1 K_2} + jn_2\sqrt{\mu_2 K_1}}$$

$$= 1 \underline{/2\tan^{-1} - \frac{n_2\sqrt{\mu_2 K_1}}{\cos\phi_1\sqrt{\mu_1 K_2}}} \tag{26}$$

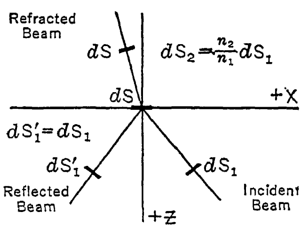


FIG. 3. For Use in Determining Energy Relationships

There is a disturbance in the second medium but no energy is transmitted, so that the wave is said to be totally reflected.

Energy Relations. To compare the energy in the incident, reflected, and refracted waves determine the rate at which energy is delivered by the incident wave to a certain area of the surface separating the two media, and the rates at which it leaves this area in the reflected and refracted waves (see Fig. 3).

The respective energies are

$$\text{Incident beam:}\quad \cos\phi_1\, dS\sqrt{\frac{K_1}{\mu_1}}\,\mathcal{E}_1^2 = \sqrt{\frac{K_1}{\mu_1}}\,\mathcal{E}_1^2 dS_1 \tag{27a}$$

$$\text{Reflected beam:}\quad \cos\phi_1\, dS\sqrt{\frac{K_1}{\mu_1}}\,\mathcal{E}'^2_1 = \sqrt{\frac{K_1}{\mu_1}}\,\mathcal{E}'^2_1 dS_1 \tag{27b}$$

$$\text{Refracted beam:}\quad \cos\phi_2\, dS\sqrt{\frac{K_2}{\mu_2}}\,\mathcal{E}_2^2 = \sqrt{\frac{K_1}{\mu_1}}\frac{\cos\phi_2}{\cos\phi_1}\,\mathcal{E}_2^2 dS_1 \tag{27c}$$

The ratio of reflected to incident energy is $\bar{\mathcal{E}}'^2_1/\bar{\mathcal{E}}_1^2$, and of refracted to the incident energy energy is $\dfrac{\cos\phi_2}{\cos\phi_1}\sqrt{\dfrac{K_2\mu_1}{K_1\mu_2}}\left(\dfrac{\bar{\mathcal{E}}_2}{\bar{\mathcal{E}}_1}\right)^2$, where the $\mathcal{E}$'s are evaluated at the interface.

Rotation of Plane of Polarization. If the incident wave is a plane-polarized wave, and the plane of polarization (which contains H_1 and hence is perpendicular to $\mathbf{\mathcal{E}}_1$) makes an angle θ_1 with the plane of incidence, then the angle between the plane of polarization and the plane of incidence of the reflected wave (θ'_1) is

$$\tan\theta_1 = -\tan\theta_1\frac{\cos(\phi_1 + \phi_2)}{\cos(\phi_1 - \phi_2)} \tag{28}$$

and for the reflected wave

$$\tan\theta_2 = \frac{\tan\theta_1}{\cos(\phi_1 - \phi_2)} \tag{29}$$

When $0 < \theta_1 < \pi/2$, the plane of polarization of the reflected wave is rotated toward the plane of incidence, while that of the refracted wave is rotated away from the plane of incidence.

CASE OF A CONDUCTING MEDIUM. If a medium is conducting, its conductivity is not zero, and by Maxwell's first equation (p. 5-50).

$$\gamma \mathbf{\varepsilon} + \frac{K\partial \mathbf{\varepsilon}}{\partial t} = \text{curl } H \tag{30}$$

where all quantities are in mks units. Assuming $\mathbf{\varepsilon}$ to be a harmonic function, it can be written $\mathbf{\varepsilon} = f\,(xyz)\ \mathbf{\varepsilon}^{j\omega t}$ and $\partial \mathbf{\varepsilon}/\partial t = j\omega \mathbf{\varepsilon}$. Hence eq. (30) becomes

$$\left(\frac{\gamma}{j\omega} + K\right)\frac{\partial \mathbf{\varepsilon}}{\partial t} = \text{curl } H \tag{31}$$

This equation is similar to that for a dielectric, the only difference mathematically being that the coefficient of $\partial \mathbf{\varepsilon}/\partial t$ is the constant K, whereas in eq. (31) it is the constant $K - j\gamma/\omega$. Hence all results obtained for insulating media are applicable to a conducting medium provided the K used in the former case is replaced by $K - j\gamma/\omega$.

If $\varepsilon = \bar{\varepsilon}\epsilon^{-\alpha x}\epsilon^{j(\omega t - \beta x)}$, which is a damped plane wave propagated in the X direction with velocity ω/β and attenuation α per unit length, then for this to be a solution of

$$-\,\omega^2\left(k\mu - j\,\frac{\mu\gamma}{\omega}\right)\varepsilon = \frac{\partial^2 \varepsilon}{\partial x^2} \tag{32}$$

(wave equation, all parameters are those of the conducting medium) it is necessary that $\alpha = 2\mu\gamma v$. For good conductors γ is large, hence α is large and the wave is highly attenuated—good conductors are poor transmitters of electromagnetic waves. Indeed, a perfect conductor ($\gamma = \infty$) will not transmit. These results hold for wavelengths not near the visible-light range when the conductor (called in this case a reflector) is an ordinary metal.

REFLECTION FROM A CONDUCTOR. If a wave traveling through an insulating medium (1) strikes the (plane) interface of an adjacent uniform conducting medium (2) all the results previously obtained for the case of two insulating media are applicable provided K_2 is replaced by $K'_2 - j\gamma/\omega$, where K'_2 now represents the dielectric constant of the conducting medium. Assuming directions shown in Fig. 2, a plane wave in the insulating medium produces in the conducting medium a wave

$$\varepsilon_2 = \bar{\varepsilon}_2\epsilon^{\alpha z}\epsilon^{j[\omega t - \beta'_2(x \sin \phi'_2 - 2 \cos \phi'_2) + \delta_2]} \tag{33}$$

where α is a positive real quantity (z is negative so the wave is damped) and ω, β'_2, ϕ'_2, and δ^2 are all real. The quantity ω is $2\pi f$ where f is the frequency of the incident wave, and v'_2, the velocity of the wave in the conducting medium, is ω/β'_2. Furthermore, $\beta'_2 = 2\pi/\lambda'_2$, where λ'_2 is the wavelength of the wave, and

$$\eta' = \frac{v_1}{v_2} = \frac{\beta'_2}{\beta_1} = \frac{\sin \phi_1}{\sin \phi'_2} \tag{34}$$

is the refractive index of medium 2 with respect to medium 1. The various quantities α, β'_2, $\sin \phi'_2$, $\cos \phi'_2$, v'_2, and η' *are all functions of the angle of incidence* ϕ_1. It is customary, when not otherwise specified, to assume that quoted values of α, β'_2, v'_2, and η' are for normal incidence ($\phi_1 = 0$). To compute these quantities for any angle of incidence, the following procedure may be used (K is written for K'_2 and subscript N refers to normal incidence; T to $2\pi/\omega$). Calculate or measure

$$\eta'_N = \frac{\sqrt{K + \sqrt{K^2 + 4\gamma^2 T^2}}}{\sqrt{2K_1}}$$

$$\alpha_N = \frac{\omega}{\sqrt{2}}\sqrt{\sqrt{K^2 + 4\gamma^2 T^2} - K} \qquad \frac{\alpha_N}{\beta_1} = \frac{\sqrt{\sqrt{K^2 + 4\gamma^2 T^2} - K}}{\sqrt{2K_1}}$$

Then η' can be found from

$$(\eta'^2 - \sin^2 \phi_1)\left(\eta'^2 - \frac{K}{K_1}\right) = \eta^2_N\left(\frac{\alpha_N}{\beta_1}\right)^2$$

and v'_2, β'_2, and ϕ'_2 follow from eq. (34). The following are of importance:

$$\eta_N = \sqrt{K'_2} \text{ only when } \gamma T \text{ is small and } K_1 = 1$$

$$\eta_N = \sqrt{\gamma T/K_1} \text{ when } \gamma T \gg 1$$

$$\alpha_N = \omega\sqrt{\eta_N^2 - K'_2}$$

$$\alpha = \eta'\beta_1 = 2\pi\eta'/\lambda \text{ when } \gamma T \gg 1$$

$$\eta_N \gg 1 \text{ and } \alpha_N = \omega\eta_N \text{ when } \gamma T \gg 1$$

The ratio of energy reflected from to energy incident upon the interface per unit time is

$$\epsilon - (2/\sqrt{\gamma T} \cos \phi_1) \text{ for } \gamma T \text{ large.}$$

As γ approaches ∞, this ratio approaches unity, showing that good conductors are good reflectors. For normal incidence ($\phi_1 = 0$) it becomes

$$1 - 2\sqrt{K_1/\gamma T}$$

and in general, assuming $\mu_1 = \mu_2$, it is

$$\frac{(n_1 - \eta' n_2)^2 + (\alpha/\beta_1)^2}{(n_1 + \eta' n_2)^2 + (\alpha/\beta_1)^2}$$

It may be noted that in a conducting medium neither the electric nor the magnetic vector lies in the wave front.

The preceding discussion has assumed K'_2 a constant, whereas it is sometimes necessary (e.g., in Heaviside layer studies) to consider K'_2 and other quantities as varying.

BIBLIOGRAPHY

Ramo, S., and J. R. Whinnery, *Fields and Waves in Modern Radio.* John Wiley (1944).
Skilling, H. H., *Fundamentals of Electric Waves.* John Wiley (1942).
Stratton, J. A., *Electromagnetic Theory.* McGraw-Hill (1941).
Schelkunoff, S. A., *Electromagnetic Waves.* D. Van Nostrand (1943).
Slater, J. C., *Microwave Transmission.* McGraw-Hill (1942).

ELECTROMECHANICAL SYSTEMS

By Knox McIlwain

Vibrations in one dimension occur frequently in systems made up of solid elements, such as a spring suspension, a pendulum, or a rocking lever. They occur occasionally in fluids, an example being the transmission of sound through a long narrow tube. Alternating electric currents in short wires will be shown to be similar analytically to unidimensional mechanical vibrations. Also devices are in common usage (microphones, speakers, etc.) which convert any of these forms of vibration into any other.

Because of the analytical similarity of the three forms of vibration, mathematical results obtained in one field may be used in solving problems in the other fields. The mathematics involved, such as the complex notation, has been applied more generally to the electrical problem than to the others, but once the analytical similarity is established advances in one field are immediately applicable to the others.

As a first approximation all the parameters are considered constants. This is reasonably true in many electrical circuits but is not generally so in mechanical systems. However, the variations for small displacements and velocities such as occur in mechanical systems used in acoustics are usually negligible. The methods of attack available when the variation of parameters must be considered are given in articles 16–24.

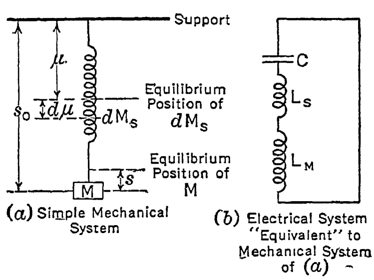


FIG. 1. Electrical and Mechanical Systems of One Degree of Freedom

DEFINITIONS. The number of independent variables required completely to specify the motion of every part of a vibrating system is a measure of the number of *degrees of freedom* of the system. When only one variable is needed the system is said to have one degree of freedom; examples of such systems are (Fig. 1): a piston moving in a cylinder, a weight hanging from a spring, cylinder rolling on a plane surface (no slipping). Systems of two or more degrees of freedom are exemplified by an automobile moving on a plane (two degrees if no skidding, three if skidding occurs), a balloon (three degrees if there is no spinning of the balloon), an airplane (six degrees if rotation is considered), a set of weights connected by springs in series (Fig. 2) which has as many degrees of freedom as there are springs; continuous vibrating systems (waves traveling along a long spring or stretched string) have an infinite number of degrees of freedom.

Consider a system of one degree of freedom (Fig. 1), and let the displacement coordinate be chosen to measure the distance from the center of the mass M to the equilibrium position. Consider that the spring has mass. The velocity of any given elementary mass of the system will then be proportional to $ds/dt(\equiv \dot{s})$, so that the kinetic energy of this elementary mass dm will be $^1/_2\, k_1\dot{s}^2 dm_1$, where k_1 is a constant dependent on the *position* of the mass. Likewise the kinetic energy of any other elementary mass dm_2 will be $^1/_2\, k_2\dot{s}^2 dm_2$, etc., and the total kinetic energy T of the system will be the sum of these, or

$$T = {}^1/_2 \int k_1\dot{s}^2\, dm_1 = {}^1/_2\, \dot{s}^2 \int k_1\, dm_1 \tag{1}$$

where the integration is to extend over the entire mass composing the system. But

$$\int k_1\, dm_1 = m_m \tag{2}$$

where m_m is a constant, called the *generalized mass*. Note that the value of m_m depends on the k's, which in turn depend on the choice of the measure of displacement. This

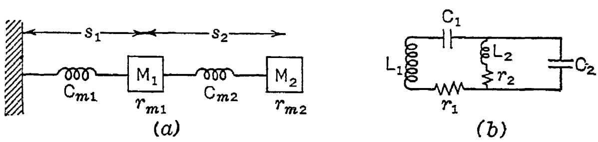


FIG. 2. Mechanical System of Two Degrees of Freedom and Electrical Equivalent

choice is largely arbitrary, and the particular one here chosen may not be the most convenient. The factors entering the problem which depend on this arbitrary choice are termed *generalized*. Thus $\dot{s}$ is a *generalized velocity* of the system; in this case it is the actual velocity of the mass M, but need not be. On the other hand, m_m is not the actual mass of the system but might be made so with a proper choice of s, the *generalized displacement*.

30. ENERGY OF MECHANICAL AND ELECTRICAL SYSTEMS

LINEAR MOTION. The kinetic energy of a mechanical system of one degree of freedom in linear motion is

$$T = {}^1/_2 m_m \dot{s}^2 \tag{3}$$

where m_m is the generalized mass of the system and s is the generalized velocity. The generalized mass can be defined as the quotient of the kinetic energy T of the system divided by the square of the generalized velocity.

The potential energy V of the system is a function of the displacement of the various parts of the system; it is independent of velocity, acceleration, etc., and in a system of one degree of freedom, since the displacements of the various parts are all proportional to the generalized displacement s, the potential energy V is a function of s only, that is

$$V = V(s) \tag{4}$$

This can be expanded in a Taylor's series:

$$V = V_0 + s\frac{\partial V}{\partial s}\bigg]_0 + \frac{s^2}{2}\frac{\partial^2 V}{\partial s^2}\bigg]_0 + \text{etc.} \tag{5}$$

where V_0 is the value of V when $s = 0$, and the subscript 0 indicates that derivatives are to be evaluated at $s = 0$. Since potential energy can be measured from any arbitrary level, V_0 may be taken as zero without loss of generality. Furthermore, in vibrating systems the generalized displacement may be so chosen that $s = 0$ for *equilibrium;* then, since $\partial V/\partial s = 0$ for equilibrium, the second term on the right-hand side of eq. (5) drops out. Making the further assumption that *terms containing higher powers of* s *than the second are negligible* in eq. (5), V reduces to

$$V = \frac{1}{2}\frac{s^2}{C_m} \tag{6}$$

where C_m is a constant $\left(\frac{1}{C_m} = \frac{\partial^2 V}{\partial s^2}\bigg]_0\right)$. C_m is called the compliance; its reciprocal, the stiffness. It will be shown later that eq. (6) is equivalent to assuming that the restoring forces of the system obey Hooke's law.

If there is dissipation (heat) in the mechanical system, *assume* that the *rate* of energy dissipation D is

$$D = r_m \dot{s}^2 \tag{7}$$

where r_m is a constant.

In many mechanical systems, particularly those concerned with the small motions usual in acoustic work, this is practically true; it is equivalent, as shown below, to assuming that the retarding frictional force is proportional to velocity, and for small motions this is approximately true of air friction. r_m is called mechanical resistance. If it is not a constant, that is, if D is not proportional to $\dot{s}^2$, but depends on higher powers of $\dot{s}$ as well, the equations cease to be linear and the methods of articles 16–20 must be used.

ROTATIONAL MOTION. The kinetic energy of a mechanical system of one degree of freedom in rotational motion is

$$T = {}^1/_2 I \dot{\phi}^2 \tag{8}$$

where I is the moment of inertia of the system, ϕ the angular displacement in radians, and $\dot{\phi}$ the angular velocity in radians per second.

The potential energy is

$$V = \frac{1}{2}\frac{\phi^2}{C_r} \tag{9}$$

where the same assumptions regarding higher powers of ϕ are made as in the case of linear motion.

Likewise the dissipation is assumed as

$$D = r_r \dot{\phi}^2 \tag{10}$$

where r_r is a constant.

ELECTRIC CURRENTS. The stored energy in an inductance L in an electrical circuit is (cf. eq. [3])

$$T = {}^1/_2 L i^2 = {}^1/_2 L \dot{q}^2 \tag{11}$$

where $i (= dq/dt \equiv \dot{q})$ is the current in the electric circuit, and q is the charge. This is often called the kinetic energy of the electrical circuit.

The energy in a condenser C is (cf. eq. [6])

$$V = \frac{1}{2}\frac{q^2}{C} \tag{12}$$

The rate of energy dissipation in an electrical circuit is (cf. eq. [7])

$$D = ri^2 \tag{13}$$

where r is the resistance of the circuit.

31. VIBRATIONS OF A SYSTEM OF ONE DEGREE OF FREEDOM

LINEAR MOTION. If a mechanical system of one degree of freedom is in a condition of stable vibration its gains and losses of energy must be equal. Hence the increase in kinetic and potential energy added to the energy dissipated must equal the work done, or

$${}^1/_2 m_m \dot{s}^2 + s^2/2C_m + \int r_m \dot{s}^2\, dt = fs \tag{14}$$

where fs, the product of force and distance, is the work done on the system. The rate of change of energy is

$$m_m \ddot{s} + r_m \dot{s} + s/C_m = f \tag{15}$$

Equation (15) is an equation in forces; that is, each term has the dimensions of a force. The first term, $m\ddot{s}$, represents the usual inertial force due to the motion of the system; the second term, $r_m\dot{s}$, represents a retarding force, proportional to the velocity of the system—this for small displacements and velocities represents approximately air friction, etc.; the third term, s/C_m, represents a force within the system proportional to the displacement. This last force is usually a restoring force, for example, the restoring force of a spring. Since this force is proportional to the displacement, Hooke's law (that the stress is equal to the strain) applies. Any internal force, such as the restoring force accompanying the straining of any part of the mechanical system, so long as the part is not stretched beyond the elastic limit, that is, so long as the restoring force is proportional to the displacement, contributes a term of the form s/C_m. On the other hand, if, instead of a restoring force, the term s/C_m were to represent a force proportional to the displacement, but tending not to restore the system to its equilibrium position but to displace the system further, then

C_m would be intrinsically negative. There is no simple analog in a passive electric circuit to this negative C_m which sometimes appears in mechanical systems. A "negative condenser" must be used, such as can be obtained in vacuum-tube circuits.

Equation (15) would be the equation of the system of Fig. 1, if the force were applied to the mass M and if s were measured from the equilibrium position. The force f is a function of t; it can be expanded in a Fourier series (see article 2) and each component of Form $F\epsilon^{j\omega t}$ treated separately. If $s = s_\epsilon^{j\omega t}$ then

$$s\left[r_m + j\omega m_m + \frac{1}{j\omega C_m}\right] = \frac{F}{j\omega} \tag{16}$$

and $s = F/j\omega z_m$, where $z_m = r_m + j\omega m_m + 1/j\omega C_m$ is called the *vector mechanical impedance* of the mechanical system. The real part of z_m is the *mechanical resistance;* the imaginary part is the *mechanical reactance.* If $\dot{s}$ were determined instead of s, assuming $\dot{s} = \dot{s}_\epsilon^{j\omega t}$, then $\dot{s} = F/z_m$. This is identical in form with the usual equation for a sine-wave current in an electric circuit.

ROTATIONAL MOTION. The above analysis can be applied to rotational motion if torque is substituted for force. Thus

$$r_r\dot{\phi} + I\frac{d\dot{\phi}}{dt} + \frac{\phi}{C_r} = L \tag{17}$$

so that $\dot{\phi} = L/z$, where $z = r_r + j[\omega I - (1/\omega C_r)]$ is called the vector rotational impedance of the system.

32. COMPARISON OF MECHANICAL AND ELECTRICAL SYSTEMS

UNITS. The analogy developed above between the equations of mechanical systems and electric circuits must not be interpreted to mean that there is any actual similarity or analogy between quantities occupying the same position in their respective equations. The inertia of the electric circuit and the mass of the mechanical system though appearing in the same place in the differential equation are otherwise quite diverse. Perhaps the difference is most convincingly shown by the fact that the analogous quantities *current* and *velocity* have different dimensions, current having the dimensions $M^{1/2}L^{1/2}T^{-1}$ in the practical electric system whereas the dimensions of velocity are LT^{-1}.

If this difference is kept in mind and it is remembered that the analogy is merely a formal one, representing the similarity of the differential equations expressing the behavior of the several systems, it is convenient to draw analogies between all quantities of the several systems. A list of these is given in Table 1. Any system of units may be used in any of the equations; only the cgs mechanical systems and the practical system for the electric circuit are shown in Table 1 on p. 5-60.

When energy equations are written in the two systems the dimensions of the equations must of course be identical, since the dimensions of energy are fixed. The usual current-electromotive force equation, however, does not have the same dimensions as a mechanical force-velocity equation. Nevertheless it is perfectly proper to set up electric circuits which are *equivalent* to mechanical systems, or vice versa, use the ordinary electric mesh equations to obtain a solution, and use this solution to specify the proper constants for either or both systems. The validity of this procedure depends on the fact that the solution of the differential equations is independent of the meaning attached to the symbols.

SYSTEMS OF MANY DEGREES OF FREEDOM. The extension of the analogy between electrical and mechanical systems to more than one degree of freedom, and the analysis of such systems which contain *both* electrical and mechanical portions, are most readily accomplished by the application of Lagrange's principle. In the application of this principle the energy equations for the whole system are first written down, in terms of measurements from an equilibrium position. Lagrange's equation that

$$\frac{d}{dt}\frac{\partial T}{\partial \dot{q}_1} - \frac{\partial(T-V)}{\partial q_1} + \frac{1}{2}\frac{\partial D}{\partial \dot{q}_1} = f_1 \tag{18}$$

is applied for each independent coordinate (each q needed in writing the energy equations).

ELECTRICAL CIRCUITS EQUIVALENT TO MECHANICAL SYSTEMS. If the energy equations of two systems can be thrown into the same form then the solutions of the two systems must be identical. The same applies to the differential equations, or to the resulting mesh equations, but in general the similarity of the systems can be established more readily by use of the energy equations.

Table 1. Analogous Quantities in Linear and Rotational Mechanical Systems and Electric Circuits

The cgs units and practical electrical units are used.

Linear Motion			Rotational Motion			Electric Circuit		
Quantity	Unit	Symbol	Quantity	Unit	Symbol	Quantity	Unit	Symbol
Force.......	dyne	f	Torque......	dyne cm	L	Electromotive force..	volt	e
Displacement	cm	s	Angular displacement	radian	ϕ	Charge.....	coulomb	q
Velocity.....	$\frac{\text{cms}}{\text{sec}}$	$\dot{s}$	Angular velocity....	$\frac{\text{radian}}{\text{sec}}$	$\dot{\phi}$	Current....	ampere	i
Mass.......	gram	m_m	Moment of inertia.....	gram cm^2	I	Inductance..	henry	L
Linear...... compliance	$\frac{\text{cms}}{\text{dyne}}$	C_m	Rotational compliance.	$\frac{\text{radians}}{\text{dyne cm}}$	C_r	Capacitance	farad	C
Mechanical resistance..	$\frac{\text{dyne sec}}{\text{cm}}$ mech.ohm	r_m	Rotational resistance...	$\frac{\text{dyne cm sec}}{\text{radian}}$ rotational ohm	r_r	Resistance..	ohm	r
Mechanical reactance..	ohm	x_m	Rotational reactance...	ohm	x_r	Reactance..	ohm	x
Mechanical impedance.	ohm	z_m	Rotational impedance..	ohm	z_r	Impedance	ohm	z
Power......	$\frac{\text{ergs}}{\text{sec}}$	P_m	Power.......	$\frac{\text{ergs}}{\text{sec}}$	P_r	Power......	watts	P

The rules for establishing the equivalence of an electric circuit with a mechanical system are therefore:

1. Write the equations for the kinetic energy, the potential energy, and the dissipation of the mechanical system.
2. There will be an electric mesh for each independent generalized coordinate.
3. Any parameter which appears in an energy equation multiplied only by the square of one generalized coordinate is a part of the corresponding electric mesh, and is not common to any other mesh.
4. Any parameter which appears multiplied by the difference of two generalized coordinates will be common to the corresponding electric meshes.
5. Any parameter which appears multiplied by the product of two generalized velocities may be represented by a mutual inductance between the corresponding electric meshes.
6. If parameters appear in the energy equations multiplied by any other combinations of the generalized coordinates, new coordinates should be chosen in an attempt to eliminate them. If this is impossible * there is no one-to-one (parameter-to-parameter) equivalent circuit.

To illustrate the method two examples will be worked out.

In the mechanical system shown in Fig. 2(a) assume that the masses M_1 and M_2 are resting on a plane surface whose frictional force is proportional to velocity. Let s_1 and s_2 be the changes in s'_1 and s'_2 from the equilibrium position. Then

$$T = \tfrac{1}{2}M_1\dot{s}_1^2 + \tfrac{1}{2}M_2(\dot{s}_1 + \dot{s}_2)^2$$

$$V = \frac{1}{2}\frac{s_1^2}{C_{m1}} + \frac{1}{2}\frac{s_2^2}{C_m{}^2}$$

and

$$D = r_{m1}\dot{s}_1^2 + r_{m2}(\dot{s}_1 + \dot{s}_2)^2$$

* Such cases will not be frequent. When encountered it is best to write force equations and attempt to obtain circuits which will represent the mechanical system by allowing electrical parameters to represent combinations of mechanical constants; for example, $L_2 = 2m_2 + m_1$ or $C_3 = C_{m1} - C_{m4} + C_{m3}$ might be required. In many such cases the utility of the method is doubtful.

The equations for T and D both contain a parameter multiplied by the *sum* of two generalized velocities squared, for which no electrical equivalent is given in the usual simple rules for independent meshes. These equations may be altered in two ways; first, the direction in which s'_2 (or s'_1) is measured may be changed; or second, new s's may be chosen. If the first alternative is employed

$$T = {}^1/_2 M_1 \dot{s}_1^2 + {}^1/_2 M_2(\dot{s}_1 - \dot{s}_2)^2$$

V is unchanged and

$$D = r_{m1}\dot{s}_1^2 + r_{m2}(\dot{s}_1 - \dot{s}_2)^2$$

The equivalent electrical circuit is shown, in which $L_1 = M_1$, $L_2 = M_2$, $r_1 = r_{m1}$, $r_2 = r_{m2}$, $C_1 = C_{m1}$ and $C_2 = C_{m2}$. That is, in the electrical circuit 1 henry of inductance is specified for each gram of mass in the mechanical system, 1 farad of capacitance for each centimeter/dyne of compliance, etc.

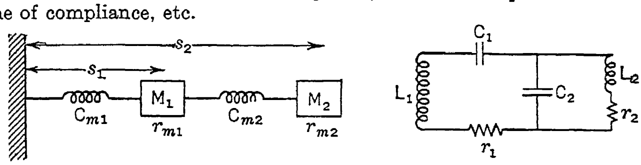


FIG. 3. Same Mechanical System as Fig. 2 and Alternate Electrical Equivalent

If the second alternative is chosen the sketch of Fig. 3 would represent the conditions. Here

$$T = {}^1/_2 M_1 \dot{s}_1^2 + {}^1/_2 M_2 \dot{s}_2^2$$

$$V = \frac{1}{2}\frac{s_1^2}{C_{m1}} + \frac{1}{2}\frac{(s_2 - s_1)^2}{C_{m2}}$$

and

$$D = r_{m1}\dot{s}_1^2 + r_{m2}\dot{s}_2^2$$

The corresponding electric circuit is shown.

It is thus possible to have more than one equivalent electric circuit for a given mechanical system, depending on the choice of independent variables for the mechanical system. Casual examination of the electric circuits reveals them as equivalent, in that although the mesh currents differ the current through any element would be the same.

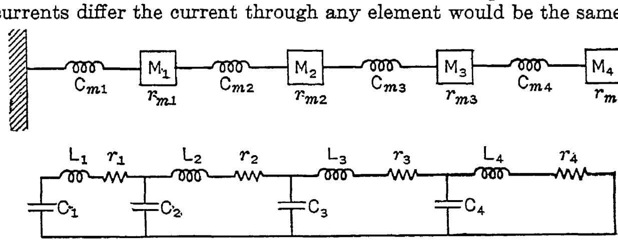


FIG. 4. Mechanical System of Four Degrees of Freedom and Electrical Low-pass Filter

In the mechanical system shown in Fig. 4 first assume that the displacements are measured between masses, as in Fig. 2. The energy equations are

$$T = {}^1/_2 M_1 \dot{s}_1 + {}^1/_2 M_2(\dot{s}_1 + \dot{s}_2)^2 + {}^1/_2 M_3(\dot{s}_1 + \dot{s}_2 + \dot{s}_3)^2 + {}^1/_2 M_4(\dot{s}_1 + \dot{s}_2 + \dot{s}_3 + \dot{s}_4)$$

$$V = \frac{1}{2}\frac{s_1^2}{C_{m1}} + \frac{1}{2}\frac{s_2^2}{C_{m2}} + \frac{1}{2}\frac{s_3^2}{C_{m3}} + \frac{1}{2}\frac{s_4^2}{C_{m4}}.$$

and

$$D = r_{m1}\dot{s}_1^2 + r_{m2}(\dot{s}_1 + \dot{s}_2)^2 + r_{m3}(\dot{s}_1 + \dot{s}_2 + \dot{s}_3)^2 + r_{m4}(\dot{s}_1 + \dot{s}_2 + \dot{s}_3 + \dot{s}_4)^2$$

In this case it is impossible to throw the equations into the usual mesh equation form by changing the direction of measurement of one or more of the variables, so the variables must be changed if the convenience of the mesh equation technique is to be utilized. If the system used in Fig. 3 is assumed

$$T = {}^1/_2 M_1 \dot{s}_1^2 + {}^1/_2 M_2 \dot{s}_2^2 + {}^1/_2 M_3 \dot{s}_3^2 + {}^1/_2 M_4 \dot{s}_4^2$$

$$V = \frac{1}{2}\frac{s_1^2}{C_{m1}} + \frac{1}{2}\frac{(s_2 - s_1)^2}{C_{m2}} + \frac{1}{2}\frac{(s_3 - s_2)^2}{C_{m3}} + \frac{1}{2}\frac{(s_4 - s_3)^2}{C_{m4}}$$

and

$$D = r_{m1}\dot{s}_1^2 + r_{m2}\dot{s}_2^2 + r_{m3}\dot{s}_3^2 + r_{m4}\dot{s}_4^2$$

The equivalent electric circuit is a low-pass filter as shown.

33. ELECTROMECHANICAL-ACOUSTIC SYSTEMS

Many vibrating systems consist of combinations of two or more of the separate energy systems discussed above. Examples of such combinations are electrically controlled vibrating reeds and tuning forks, microphones and speakers of all varieties, fluid-type automobile stabilizers, and almost all musical instruments.

In such systems energy is converted from one form to another within the system. The methods of converting energy are many, so that the forms of the interaction between the different portions of the system are many, but the same general method of determining the force equations that was used for homogeneous systems can be applied.

The energy of each portion of the system should be set up exactly as for a homogeneous system in terms of the constants of the particular portion. Some of these constants will be functions of the velocity or displacement, etc., of some other portion of the system, and these must be so specified. Then the "force" equations may be obtained directly by means of Lagrange's equation, and from these the mesh equations may be written. These mesh equations will not usually be of the same dimensions, but this fact may be disregarded since the interaction factors between the equations will be such that when the equations are solved simultaneously these differences will be automatically compensated. One set of self-consistent units is shown in Table 2. Several particular cases will be developed to show the operation of the method.

ELECTROMECHANICAL SYSTEMS, ELECTROSTATICALLY COUPLED. Consider a system consisting of two metal plates, one fixed and immovable, the other held in position by a spring (cf. Fig. 5) and resting on a rough surface. An electric circuit connected between the two plates contains an inductance, a resistance, and a source of voltage.

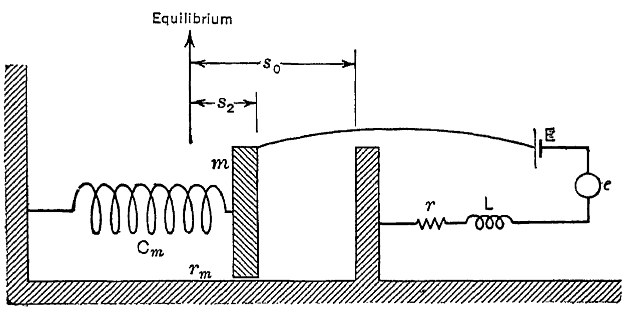


FIG. 5. Electromechanical System, Electrostatically Coupled

The potential energy V_m stored in the spring (measuring energy from its value at the equilibrium position, when the electric charge is zero) is

$$V_m = \frac{1}{2}\frac{s_2^2}{C_m}$$

where C_m is the compliance of the spring and s_2 is the distance the spring is stretched by moving the movable plate toward the fixed plate.

The potential energy V_e stored in the condenser is

$$V_e = \frac{1}{2}\frac{q^2}{C} = \frac{1}{2}\frac{q^2}{C'}(s_0 - s_2)$$

where q is the charge on the condenser and C its capacitance. The value of C is

$$C = \frac{kS}{4\pi(s_0 - s_2)} = \frac{C'}{s_0 - s_2}$$

where k is the dielectric constant, S the cross-section of the plates, and $(s_0 - s)$ the distance between them.

The kinetic energy of the system is

$$T = \frac{m\dot{s}_2^2}{2} + \frac{Li^2}{2}$$

where m is the generalized mass of the movable plate and spring together and L is the inductance of the electrical system. The dissipation of the system is

$$D = r_m\dot{s}_2^2 + ri^2$$

Table 2. A Self-consistent Set of Units for Use in Electromechanical-acoustic Systems *

Linear Motion			Rotational Motion			Electric Circuit			Acoustic System		
Quantity	Unit	Symbol	Quantity	Unit	Symbol	Quantity	Unit	Symbol	Quantity	Unit	Symbol
Displacement..	cm	s	Angular displacement.	radian	ϕ	Charge......	coulomb	q	Displacement....	cm	s
Velocity.......	cm sec	$\dot{s}$	Angular velocity.....	$\frac{\text{radian}}{\text{sec}}$	$\dot{\phi}$	Current......	ampere	i	Particle velocity.	cm sec	$\dot{s}$
Force.........	dynes $\times 10^{-7}$	f_m	Torque.......	dyne cm $\times 10^{-7}$	L	Voltage......	volt	e	Force (pressure times cross-section).......	dynes $\times 10^{-7}$	f_f
Mass.........	grams $\times 10^{-7}$	m_m	Moment of inertia	gram cm$^2 \times 10^{-7}$	I	Inductance...	henry	L	Mass..........	grams $\times 10^{-7}$	m_f
Linear compliance...	$\frac{\text{cm} \times 10^7}{\text{dynes}}$	C_m	Rotational compliance..	$\frac{\text{radian}}{\text{dyne cm}} \times 10^7$	C_r	Capacitance..	farad	C	Fluid compliance....	$\frac{\text{cm} \times 10^7}{\text{dynes}}$	C_f
Mechanical resistance....	$\frac{\text{dyne sec}}{\text{cm}} \times 10^{-7}$	r_m	Rotational resistance....	$\frac{\text{dyne cm sec}}{\text{radian}} \times 10^{-7}$	r_r	Resistance...	ohm	r_e	Acoustic resistance......	$\frac{\text{dyne sec}}{\text{cm}} \times 10^{-7}$	r_f
Mechanical reactance.....	mech ohm † $\times 10^{-7}$	x_m	Rotational reactance....	rot. ohm † $\times 10^{-7}$	x_r	Reactance....	ohm	x_e	Acoustic reactance......	acoust. ohm † $\times 10^{-7}$	x_f
Mechanical impedance...	mech. ohm † $\times 10^{-7}$	z_m	Rotational impedance...	rot. ohm † $\times 10^{-7}$	z_r	Impedance...	ohm	z_e	Acoustic impedance.....	acoust. ohm † $\times 10^{-7}$	z_f
Power.........	watts	P_m	Power.......	watts	P_r	Power.......	watts	P_e	Power..........	watts	P_f

* Based on the practical electrical system and the centimeter. An alternate set of units is given in Olsen and Massa, Applied Acoustics.
† A mechanical, or an acoustical, ohm is one dyne second per centimeter. A rotational ohm is one dyne centimeter second per radian.

The force equations may be obtained by applying Lagrange's equations to these energy equations; they are

$$m\ddot{s}_2 + \frac{s_2}{C_m} - \frac{q^2}{2C'} + r_m\dot{s}_2 = f$$

and

$$L\frac{di}{dt} + \frac{s_0}{C'}q - \frac{q}{C'}s_2 + ri = e$$

Assuming that q can be split into a steady or average part q_0 and a varying part q_1 and that q_1 is considerably smaller than q_0, so that squares of q_1 are negligible,

$$-\frac{q^2}{2C'} = -\frac{(q_0^2 + 2q_0q_1 + q_1^2)}{2C'}$$

The first term on the right-hand side of this equation represents a steady pull or initial set of the plunger accompanying the initial charge on the condenser. The second term reduces to $\frac{q_0}{C'}q_1$, and the third term is negligible by assumption. Neglecting the steady terms the force equations become

$$m\ddot{s}_2 + \frac{s_2}{C_m} - \frac{q_0}{C'}q_1 + r_m\dot{s}_2 = f_2 \tag{19a}$$

$$L\frac{di_1}{dt} + \frac{q_0}{C'}q_1 - \frac{q_0}{C'}s_2 + ri_1 = e_1 \tag{19b}$$

which for steady-state sine waves may be written in the familiar mesh equation form

$$F_2 = z_{22}\dot{S}_2 + z_{12}I_1 \tag{20a}$$

$$E_1 = z_{12}\dot{S}_2 + z_{11}I_1 \tag{20b}$$

where mesh 2 represents the mechanical and mesh 1 the electrical system. z_{11} and z_{22} are the usual mechanical and electrical impedances and F_2 and E_1 the respective "forces." $z_{12} = -q_0/j\omega C' = -1/j\omega C_{12}$ is the interaction factor. The dimensions of $1/C_{12}$ are such that multiplication of it by a charge gives a quantity with the dimensions of a force, while multiplication of it by a distance or length produces a quantity with the dimensions of an electromotive force.* It is energy divided by distance and charge, and its use is valid for any system of electrical units in combination with any system of mechanical units provided those units were used in evaluating z_{12}.

Equations (20) are then in proper form to use for any case of mechanical and electrical systems where the interaction between the two systems depends on varying the capacitance of the electrical circuit by the motion of a part of the mechanical system.

ELECTROMECHANICAL SYSTEMS, MAGNETICALLY COUPLED. When the interaction between the electric and mechanical systems occurs through magnetic attrac-

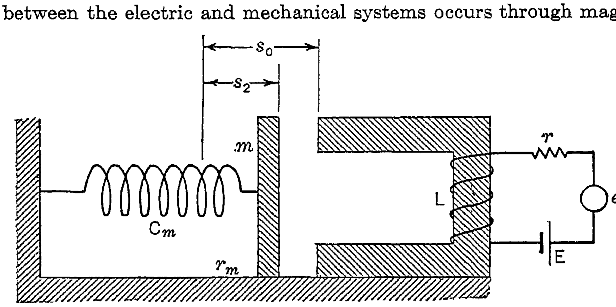


FIG. 6. Electromechanical System, Electromagnetically Coupled

tion the inductance of the electric circuit is varied by the mechanical motion. The energy equations for the system of Fig. 6 are (s_0 is the equilibrium position)

$$V = \frac{\dot{s}_2^2}{2C_m}$$

$$T = \frac{m\dot{s}_2^2}{2} + \frac{Li^2}{2}$$

and

$$D = r_m\dot{s}_2^2 + ri^2$$

* The factor $1/C_{12}$ has the dimensions of an electric field, although it is a rather unusual expression for such a field.

Here L, the self-inductance of the electric circuit, is a function of the displacement s_2. Since most of the reluctance of the magnetic circuit is in the air gap rather than in the iron portions, L can be assumed to vary as the first power of s for small displacements, with little error. Hence

$$L = L_0(1 + bs_2)$$

where b is a constant depending on the configuration of the physical system and on the systems of units used.

Applying Lagrange's equations, the force equations for the system are

$$m\ddot{s}_2 - \frac{bL_0 i^2}{2} + \frac{s_2}{C_m} r_m \dot{s}_2 = f$$

and

$$L\frac{di}{dt} + bL_0 i\dot{s}_2 + \frac{q}{C} + ri = e + E$$

Assuming that i is composed of a steady, or average, value i_0 and a varying part i_1, and that i_1 is considerably smaller than i_0, so that squares of i_1 are negligible,

$$-\frac{bL_0 i^2}{2} = -\frac{bL_0(i_0^2 + 2i_0 i_1 + i_1^2)}{2}$$

The first term on the right-hand side again represents a steady pull or initial set of the plunger accompanying the average or d-c exciting current of the magnet. The second term reduces to $-(bL_0 i_0)i_1$, and the third term is negligible by assumption. Disregarding the steady terms the force equations become

$$m\ddot{s}_2 + \frac{s_2}{C_m} - bL_0 i_0 i_1 + r_m \dot{s}_2 = f_2 \tag{21a}$$

and

$$L\frac{di_1}{dt} + \frac{q_1}{C} + bL_0 i_0 \dot{s}_2 + ri_1 = e_1 \tag{21b}$$

which for steady-state sine waves may be written

$$E_1 = z_{11}I_1 + z_{12}\dot{s}_2 \tag{22a}$$

$$F_2 = z_{21}I_1 + z_{22}\dot{s}_2 = -z_{12}I_1 + z_{22}\dot{s}_2 \tag{22b}$$

where mesh 1 represents the electrical and mesh 2 the mechanical system, E_1 and F_2 represent the respective "forces." *In this case, however, the interaction factors are not equal*, but $z_{12} = -z_{21}$.

The solutions for current and velocity give

$$I_1 = \frac{E_1 z_{22} - F_2 z_{12}}{z_{11}z_{22} + z_{12}^2}$$

and

$$\dot{s}_2 = \frac{F_2 z_{11} + E_1 z_{12}}{z_{11}z_{22} + z_{12}^2}$$

indicating that, when a mechanical force is applied to the plunger in phase with an applied electromotive force, the velocity of the plunger is increased but the electric current is decreased. Analysis of the physical action also indicates this result since increase in the electromotive force attracts the plunger, but the inward motion of the plunger due to an applied force increases the inductance of the electric circuit and so decreases the current flow.

These are the equations of a simple moving-armature telephone receiver such as that shown in Fig. 7, when eddy currents are neglected. The *input impedance* (z_i) of the receiver is the ratio of E_1 to I_1, when there is no externally applied force (that is, when $F_2 = 0$), hence

$$z_i = z_{11} + \frac{z_{12}^2}{z_{22}}$$

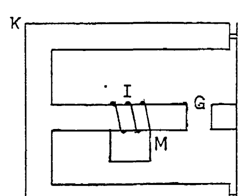


FIG. 7. Simple Telephone Receiver (No Eddy Currents)

The *blocked impedance* of a receiver is defined as the impedance measured when the diaphragm is constrained from moving (diaphragm held at $s = 0$). The *motional impedance* is the difference between z_i and the damped impedance; that is, it is that part of z_i due to the motion of the diaphragm. Evidently for the simple receiver the damped impedance is z_{11} (which is the impedance of the electric circuit considered alone) and the motional impedance $z_{\text{mot.}}$ is

$$z_{\text{mot.}} = \frac{z_{12}^2}{z_{22}} \tag{23}$$

Since z_{12} is a constant, independent of frequency by definition, it follows that, if the real part of $z_{\text{mot.}}$ (the resistance component of $z_{\text{mot.}}$) and corresponding values of the imaginary component (reactance component) be plotted for various frequencies, the resulting graph will be a circle, called the *motional impedance* circle, tangent to the j axis, with center on the positive real axis (see *High-frequency Alternating Currents*, McIlwain and Brainerd, John Wiley & Sons, Chapter XIII).

MECHANICAL-ACOUSTIC SYSTEM. The simplest mechanical-fluid system is that where a plunger moves in a fluid displacing a volume of the fluid (see Fig. 8).

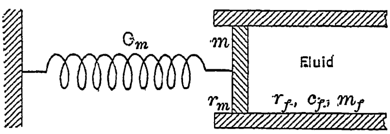


FIG. 8. Simple Mechanical-acoustic System

The potential energy of a volume of fluid moving as a unit is

$$T_f = {}^1/_2 \rho l S \dot{s}^2 = \frac{m_f}{2} \dot{s}^2$$

where ρ is the density, l the length, S the cross-section, and m_f the mass of the fluid, and $\dot{s}$ is the particle velocity. If the fluid is incompressible, $\dot{s}$ for the fluid will equal $\dot{s}$, the velocity of the plunger, and m_f will be the actual mass of the fluid; if the fluid is compressible $\dot{s}$ may still be taken as the velocity of the plunger but the generalized mass of the fluid will be less than the actual mass. The potential energy and dissipation are

$$V_f = \frac{s^2}{2C_f}$$

and

$$D_f = r_f \dot{s}^2$$

The total energy equations of the system are

$$V = \frac{s^2}{2C_m} + \frac{s^2}{2C_f}$$

$$T = \frac{m_m \dot{s}^2}{2} + \frac{m_f s^2}{2}$$

and

$$D = r_m \dot{s}^2 + r_f \dot{s}^2$$

where r_f and C_f as well as m_f are generalized parameters. Operating with Lagrange's equation there results

$$f = \left(\frac{1}{C_m} + \frac{1}{C_f}\right) s + (m_m + m_f)\ddot{s} + (r_m + r_f)\dot{s} \tag{24}$$

Thus this combination of a mechanical and an acoustic system can be represented by one equation with only one coordinate, instead of the two equations which might have been expected. Each parameter of the mechanical system has been altered by the presence of the acoustic chamber, and the resonance frequency has been lowered.

Assuming sine-wave excitation, eq. (24) becomes

$$F_m - F_f = \left[(r_m + r_f) + j\left(\omega m_m + \omega m_f - \frac{1}{\omega C_m} - \omega C_f\right)\right]\dot{s}$$

or

$$F_m - F_f = z_{mf}\dot{s}$$

where

$$z_{mf} = \sqrt{(r_m + r_f)^2 + \left(\omega m_m + \omega m_f - \frac{C_m + C_f}{\omega C_f C_m}\right)^2}$$

and

$$\theta_{mf} = \tan^{-1} \frac{\omega m_m + \omega m_f - \dfrac{C_m + C_f}{\omega C_f C_m}}{r_m + r_f}$$

The net result of the air chamber is thus to increase the resistance and inertial effect of the plunger but to decrease its compliance. In general both the magnitude and phase angle of the impedance are altered.

BIBLIOGRAPHY

McIlwain, K., and J. G. Brainerd, *High-frequency Alternating Currents.* John Wiley (1939).
Olson, H. F., and F. Massa, *Applied Acoustics.* Blakiston (1934).
den Hartog, J. P., *Mechanical Vibrations.* McGraw-Hill (1934).
Crandall, I. B., *Theory of Vibrating Systems and Sound.* Van Nostrand (1927).
Timoshenko, S., *Vibration Problems in Engineering.* Van Nostrand (1928).
Mason, W. P., *Electromechanical Transducers and Wave Filters.* Van Nostrand (1942).

Articles

Wegal, R. L., Theory of Telephone Receivers, *J. A.I.E.E.*, Vol. 40, 791 (October 1921).
Hussey, L. W., and L. R. Wrathall, Oscillations in an Electromechanical System, *B.S.T.J.*, Vol. 15, 441 (July 1936).
Mason, W. P., and R. A. Sykes, Electrical Wave Filters Employing Crystals with Normal and Divided Electrodes, *B.S.T.J.*, Vol. 19 (April 1940).
Mason, W. P., Electrical and Mechanical Analogies, *B.S.T.J.*, Vol. 20, 405 (October 1941).

SECTION 6

PASSIVE CIRCUIT ELEMENTS

PASSIVE CIRCUIT ELEMENTS

SINGLE-MESH AND COUPLED CIRCUITS

By Vernon D. Landon and Knox McIlwain

The individual elements used in communication circuits must be carefully designed to meet certain definite specifications if the circuit as a whole is to function properly. Among the more important factors to be considered are the efficiency and load-carrying capacity of the element, its impedance, its selectivity, and the introduction of the various forms of distortion.

Circuit requirements frequently demand not only that the transmission (ratio of output to input) have a particular value at a given frequency, but also that the variation of transmission with frequency, called the transmission-frequency characteristic, have a definite form.

1. SERIES RESONANT CIRCUITS

Figure 1 represents a series circuit in which the resistance of all the elements of the circuit is lumped into one resistance r, and similarly for the inductance and capacitance. For a given value of impressed voltage the current in this circuit depends on r, L, and C, and the frequency of the impressed voltage. For given values of L and C there will be one frequency, and only one, for which $\omega L = 1/\omega C$. At this frequency (f_r), which is called the *resonant frequency* of the circuit, the current is maximum for any given value of voltage and is in phase with the impressed voltage. The resonant frequency is given by

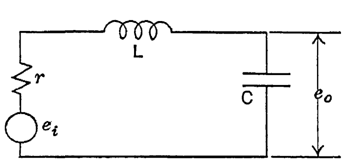


FIG. 1. Simple Series Resonant Circuit

$$f_r = \frac{1}{2\pi\sqrt{LC}} \tag{1}$$

where L is in henrys and C in farads. When L is expressed in millihenrys and C in microfarads eq. (1) becomes:

$$f_r = \frac{5033}{\sqrt{LC}} \tag{1a}$$

The absolute value of the current I_r in the circuit at this frequency is

$$I_r = \frac{E}{z_r} = \frac{E}{r} \tag{2}$$

For frequencies less than f_r, $\omega L < \dfrac{1}{\omega C}$ and the current will lead the impressed voltage (β is negative), whereas for frequencies greater than f_r the current will lag the voltage (β is positive, see Section 5, article 2).

In Fig. 2 are shown graphs of the absolute values of current versus frequency in a simple series circuit for two values of resistance.

VOLTAGE RELATIONS. At the resonant frequency the magnitude of the impressed voltage is given by $E = rI$, and the voltages across the inductor and condenser are $E_L = 2\pi f_r LI$ and $E_c = \dfrac{I}{2\pi f_r C}$. Since the only condition necessary for resonance is that the sum of the reactances is zero, either reactance may be many times as great as the circuit resistance; thus the voltage across the inductor, or the condenser, may be several hundred times the impressed voltage. Figure 3 shows how the voltage components in a circuit vary as the frequency is changed.

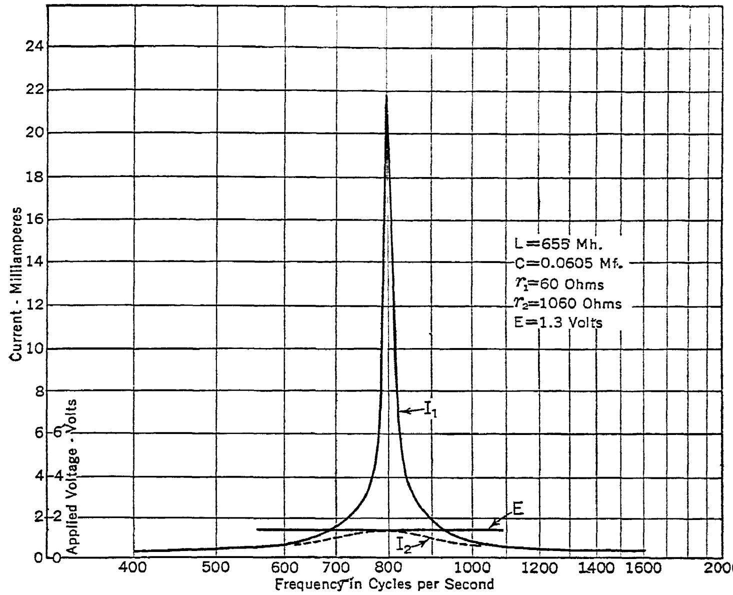


FIG. 2. Variation of Current with Frequency in a Simple Series Circuit (I_1 is for resistance r_1, and I_2 for r_2)

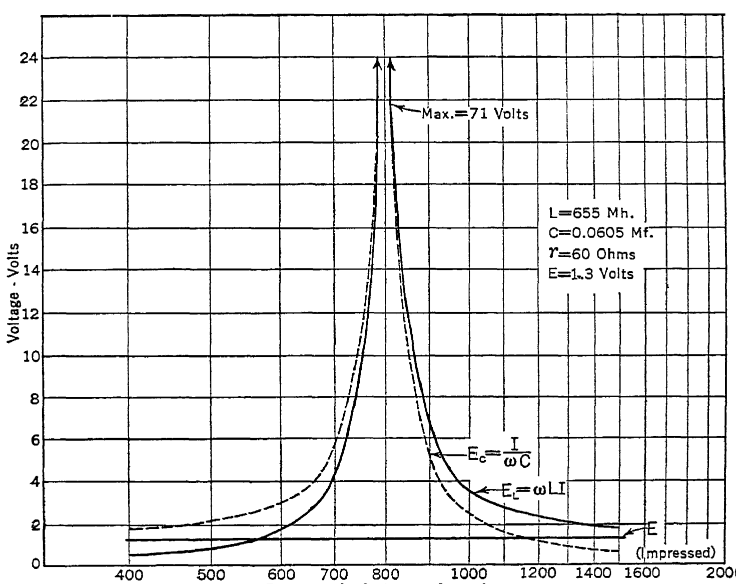


FIG. 3. Variation of Voltage Components with Frequency in a Simple Series Circuit

CIRCUIT Q. The per cent change in current from its maximum value for a given percentage frequency change depends on the relative values of r, L, and C. As a measure of this change a factor Q is defined as $\omega L/r$. From this the voltage across the coil (or condenser) of a simple series circuit at resonance is

$$E_L = \frac{2\pi f_r LE}{r} = EQ \tag{3}$$

For a given value of inductance the circuit having the higher Q will have smaller r, will have higher resonance current in comparison to current off resonance, will be represented by a more sharply peaked current vs. frequency curve, and is said to be more *sharply tuned.*

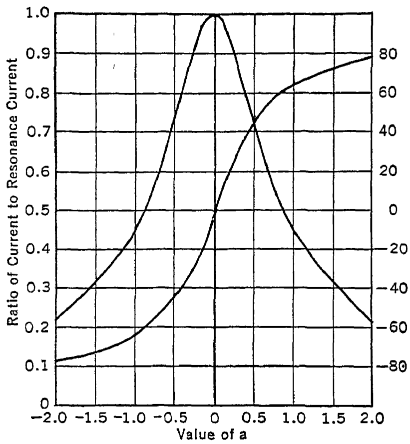


FIG. 4. Values of a Universal Resonance Curve. Calculated from

$$\frac{1}{1 + \frac{a}{Q} + ja\left(\dfrac{2 + \frac{a}{Q}}{1 + \frac{a}{Q}}\right)} \doteq \frac{1}{1 + j2a}$$

since a/Q is quite small. For parallel circuits change ratio of currents to ratios of impedances and leading phase angles to lagging.

Universal Resonance Curve. Terman has shown that, when Q is constant (loss resistance proportional to frequency), the relation of resonant current to current off resonance for any circuit can be plotted in a *universal resonance* curve as shown in Fig. 4. If the ratio of the actual current and the current at resonance are known, the ordinate is established and therefore the two possible values of a. From these the required deviation in frequency is determined from the relation

$$a = Q\,\frac{\text{Cycles off resonance}}{\text{Resonant frequency}} \tag{4}$$

Conversely, if the deviation in frequency is known, a can be readily found and thence the relative response.

2. PARALLEL RESONANT CIRCUITS

It is not so easy to define resonance in a parallel circuit. If the condition $x_1 = \omega L = -x_2 = 1/\omega C$ is used, there results

$$r = \frac{r_1 r_2 + x_1^2}{r_1 + r_2} \tag{5}$$

and

$$x = \frac{r_2 - r_1}{r_1 + r_2}\,x_1 \tag{6}$$

so that as r_1 and r_2 decrease r rises indefinitely and theoretically becomes infinite when the circuit resistance is zero. However, x is not zero at the frequency f_r, so that the current is not in phase with the voltage, nor is z a maximum. Either of these last conditions may be the important one, and the condition of z maximum (I minimum) is certainly the easiest to measure. If $r_1 = r_2$ the three conditions coincide, and are but slightly different when r_1 and r_2 are small.

In the particular important case of a coil and condenser in parallel (Fig. 5) these equations reduce (neglecting the resistance of the condenser) to

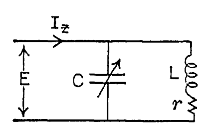


FIG. 5. Simple Parallel Resonant Circuit

$$r = \frac{L}{r_L C} \quad \text{and} \quad x = -\sqrt{\frac{L}{C}} \tag{7}$$

and the frequency for which $x = 0$ is

$$F_{(x=0)} = \frac{1}{2\pi}\sqrt{\frac{1}{LC}\,\frac{(L - r_L^2 C)}{L}} \tag{8}$$

Graphs of r, x, and **Z** for such a circuit are shown in Fig. 6.

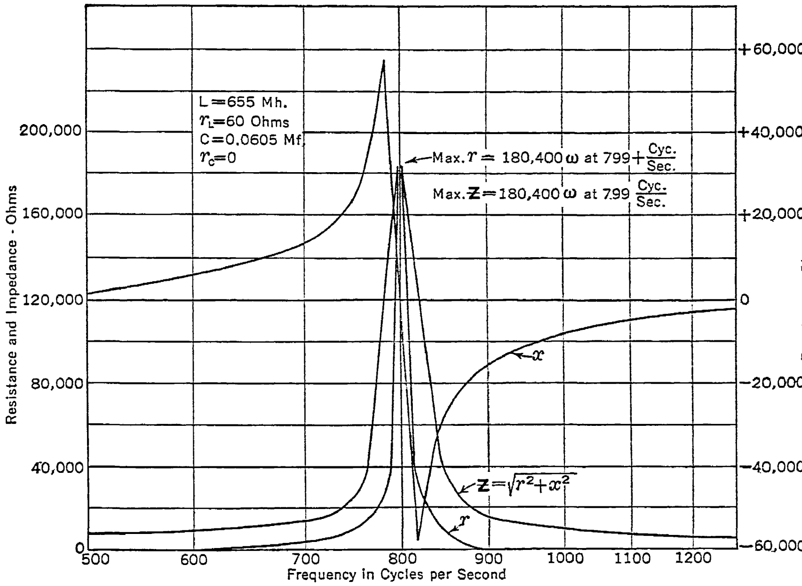


FIG. 6. Variation of the Impedance of a Simple Parallel Circuit with Frequency

Another case of particular interest exists when $r_L = r_C = \sqrt{L/C}$. Then for any frequency

$$r = \sqrt{L/C} = r_L = r_C \quad \text{and} \quad x = 0 \tag{9}$$

Thus the apparent resistance is constant with frequency, and the apparent reactance is zero for all frequencies.

TUNING. Although they are subject to the same limitations as regards controlling selectivity and frequency characteristic as the series resonant meshes, simple parallel tuned circuits are frequently used as coupling elements between vacuum tubes. When the circuit of Fig. 5 is adjusted for unity power factor, $I_c = Ex_L/(r^2 + x_L^2)$, $I_L = E/\sqrt{r^2 + x_L^2}$, $I_z = ErC/L$, and $z = L/rC$. If the capacitance only is varied, unity power factor coincides with the condition of minimum line current and with maximum impedance.

When the circuit is tuned by adjustment of the inductance this is not true, and in this case full output is obtained for a slight detuning from the condition of minimum plate current. If the resistance remains constant the maximum impedance is

$$z = \frac{2x_C r}{\sqrt{4r^2 + x_C^2} - x_C}$$

where

$$x_L = r\sqrt{\frac{2\sqrt{4r^2 + x_C^2}}{\sqrt{4r^2 + x_C^2} - x_C} - 1}$$

whereas if the phase angle (x_L/r) of the inductance remains constant the maximum impedance is $z = \sqrt{r^2 + x_L^2}\, x_L/r$ which occurs when $x_L = x_C$.

(For further details see R. Lee, *Proc. I.R.E.*, Vol. 21, 271.)

3. ATTENUATORS, PADS

Use of the concepts and equations of insertion loss (Section 5, article 7) permits the designs of definite elements (either dipoles or quadripoles) which may be inserted into circuits to produce definite effects at a given frequency. Such elements are called *attenuator sections*, or sometimes *pads;* more complicated sections which aim to provide a desired definite variation of loss with frequency are termed *distortion correctors*, or equalizers.

MATCHED IMPEDANCES. The simplest design occurs when it is desirable to have uniform loss at all frequencies, in which case the image impedances of the network must match those of the circuit into which it is inserted. By Thévenin's theorem (p. 5-12) the circuit, no matter how complicated, can be reduced to a generator in series with a simple impedance on each side of the point of insertion. (See Figs. 7 and 8.)

These impedances are designated by Z and z and their ratio by $Z/z = s^2$. Either T or π type sections may be utilized. The formulas are most useful for resistive networks.

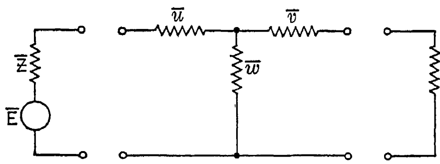


FIG. 7. T Section Attenuator

T SECTIONS. To design a T section (see Fig. 7) which causes a loss of n decibels, read off the current ratio (k) corresponding to such a loss in the decibel table, Section 1. Then the proper values for the arms of the T section are

$$u = \left(\frac{1 + k^2 - 2k/s}{1 - k^2}\right) Z, \quad v = \left(\frac{1 + k^2 - 2ks}{1 - k^2}\right) z, \quad \text{and} \quad w = 2\left(\frac{k}{1 - k^2}\right) sz$$

When the impedances in either direction are equal,

$$Z = z, \quad \text{then} \quad s = 1 \quad \text{and} \quad u = v = \frac{1 - k}{1 + k} z \quad \text{and} \quad w = \frac{2k}{1 - k^2} z$$

For instance, to cause a loss of 3 db in a circuit where $Z = z = 500 \angle 60°$, $k = 0.7080$, $u = v = 85.5 \angle 60°$, and $w = 1420 \angle 60°$.

To design a balanced T or H section simply put $\frac{1}{2}$ u and $\frac{1}{2}$ v in each series leg.

π SECTIONS. The proper values for the arms of the π section of Fig. 8 are

$$u = \left(\frac{1 - k^2}{1 - 2ks + k^2}\right) Z, \quad v = \left(\frac{1 - k^2}{k}\right) \frac{sz}{2}, \quad \text{and} \quad w = \left(\frac{1 - k^2}{1 - \frac{2k}{s} + k^2}\right) z$$

When the impedances in either direction are equal, $Z = z$, then $s = 1$ and $u = w = \left(\frac{1 + k}{1 - k}\right) z$ and $v = \frac{z}{2}\left(\frac{1 - k^2}{k}\right)$

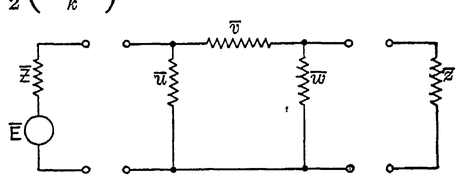


FIG. 8. π Section Attenuator

Tabular aids in computing such sections and rules for other types of sections are given by McElroy (*Proc. I.R.E.*, Vol. 23, 213 [March 1935]).

NON-MATCHED IMPEDANCES. When the impedance variation with frequency of the terminating circuits is such that it cannot be duplicated in a simple network, the insertion loss of the network at any frequency can be obtained by the methods of Section 5, article 8. Exact methods of design of sections to insure a predetermined variation of loss (and phase change) with frequency are given by Mead (*B.S.T.J.*, Vol. 7, 195), Zobel (*B.S.T.J.*, Vol. 7, 438) and Gewertz (*Network Synthesis*, Williams and Wilkins Co. [1933]).

4. COUPLED CIRCUITS

Two electric meshes are said to be coupled to each other when they have an impedance in common, so that a current in one causes a voltage in the other. The common, or *mutual*, impedance is defined as the factor by which the current in one mesh must be multiplied to give the voltage, due to that current, in the second mesh. It may be a pure resistance, or a pure reactance, in which case the meshes are said to have a *pure* coupling; or the common impedance may be complex, in which case the coupling is said to be *complex.*

Figures 9 and 10 illustrate some of the more usual types of coupling, although a complicated intervening network may be considered as simply a coupling impedance if the relation between current in one mesh and resulting voltage in the other is of interest.

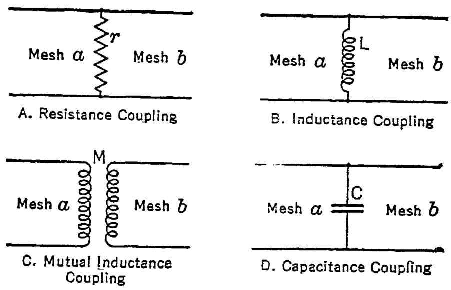


FIG. 9. Simple Types of Coupling

MUTUAL IMPEDANCES. The mutual impedances of the circuits shown in Fig. 9 are $Z_M = r, j\omega L, j\omega M, 1/j\omega C$. In Fig. 10, case E, the mutual impedance is $j\omega(L' + M)$ if the windings are in the same sense and $j\omega(L' - M)$ if they are in opposite sense. In case F, $Z_M = j(\omega L - 1/\omega C)$; in case G, $Z_M = j(\omega L/\omega^2 LC - 1)$. When the middle mesh of case H is considered as a mutual impedance between meshes a and b its value is

$$j\left[\frac{\omega^2 L_1 L_2 - \omega^2 M^2 - (M/C)}{\omega(L_1 + L_2 - 2M) - (1/\omega C)}\right]$$

when the transformer windings are so connected as to have minimum voltage across the coupling condenser. Reversing one winding reverses the sign of M.

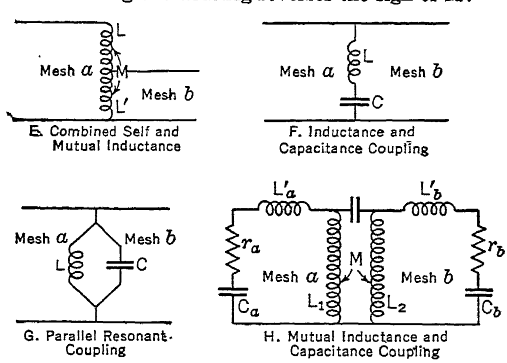


FIG. 10. Common Types of Involved Coupling

5. CURRENTS AND VOLTAGES IN COUPLED CIRCUITS

If two meshes are coupled as shown in Fig. 11 the primary current is

$$\mathbf{I}_a = \mathbf{E}/(r_a' + jx_a') = \mathbf{E}/\mathbf{z}_a' \tag{10}$$

where

$$r_a' = r_{aa} + \frac{x_m^2 r_{bb} - r_m^2 r_{bb} - 2r_m x_m x_{bb}}{z^2_{bb}} \tag{10a}$$

and

$$x_a' = x_{aa} + \frac{r_m^2 x_{bb} - x_m^2 x_{bb} - 2r_m x_m r_{bb}}{z_{bb}^2} \tag{10b}$$

Note that in these equations r_{aa} and r_{bb} are the *self-resistances* of the respective meshes ($r_{aa} = r_{ao} + r_m$), and x_{aa} and x_{bb} are the *self-reactances.* (For a complete discussion of self-impedances and mesh currents see Section 5, article 4. z_a' is called the *equivalent primary impedance* and r_a' and x_a' the *equivalent primary resistance* and *reactance.* The

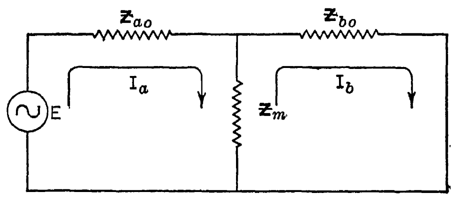


FIG. 11. Two Meshes Coupled through an Impedance $\mathbf{z}_m$

second term on the right of eq. (10*a*) is called the *transferred resistance*, and the corresponding term of eq. (10*b*) the *transferred reactance.*

The secondary current is

$$\mathbf{I}_b = \mathbf{z}_m\mathbf{I}_a/\mathbf{z}_{bb} = \mathbf{z}_m\mathbf{E}/\mathbf{z}_{aa}\mathbf{z}_b' \tag{11}$$

where

$$r_b' = r_{bb} + \frac{x_m^2 r_{aa} - r_m^2 r_{aa} - 2r_m x_m x_{aa}}{z_{aa}^2} \tag{11a}$$

and

$$x_b' = x_{bb} + \frac{r_m^2 x_{aa} - x_m^2 x_{aa} - 2r_m x_m r_{aa}}{z_{aa}^2} \tag{11b}$$

RESISTANCE COUPLING. When the coupling is a pure resistance, eqs. (11*a* and *b*) reduce to

$$r_b' = r_{bb} - r_m^2 r_{aa}/z_{aa}^2 \tag{12a}$$

and

$$x_b' = x_{bb} + r_m^2 x_{aa}/z_{aa}^2 \tag{12b}$$

so that the secondary current is

$$\mathbf{I}_b = \frac{r_m\mathbf{E}}{\mathbf{z}_{aa}\left[r_{bb} - \dfrac{r_m^2 r_{aa}}{\mathbf{z}_{aa}^2} + j\left(x_{bb} + \dfrac{r_m^2 x_{aa}}{z_{aa}^2}\right)\right]} \tag{13}$$

and its absolute value is

$$I_b = \frac{r_m E}{\sqrt{r_{aa}^2 + x_{aa}^2}\sqrt{\left(r_{bb} - \dfrac{r_m^2 r_{aa}}{z_{aa}^2}\right)^2 + \left(x_{bb} + \dfrac{r_m^2 x_{aa}}{z_{aa}^2}\right)^2}} \tag{13a}$$

The maximum value possible for I_b (for optimum tuning arrangements, $x_{aa} = x_{bb} = 0$) is

$$I_{b\,\mathrm{max\,max}} = \frac{r_m E}{r_{aa} r_{bb} - r_m^2} \tag{14}$$

REACTANCE COUPLING. When the coupling is a pure reactance, eqs. (11*a* and *b*) reduce to

$$r_b' = r_{bb} + x_m^2 r_{aa}/z_{aa}^2 \tag{15a}$$

and

$$x_b' = x_{bb} - x_m^2 x_{aa}/z_{aa}^2 \tag{15b}$$

so that the expression for the secondary current becomes

$$\mathbf{I}_b = \frac{jx_m\mathbf{E}}{\mathbf{z}_{aa}\left[r_{bb} + \dfrac{x_m^2 r_{aa}}{\mathbf{z}_{aa}^2} + j\left(x_{bb} - \dfrac{x_m^2 x_{aa}}{z_{aa}^2}\right)\right]} \tag{16}$$

and its absolute value is

$$I_b = \frac{x_m E}{\sqrt{r_{aa}^2 + x_{aa}^2}\sqrt{\left(r_{bb} + \dfrac{x_m^2 r_{aa}}{z_{aa}^2}\right)^2 + \left(x_{bb} - \dfrac{x_m^2 x_{aa}}{z_{aa}^2}\right)^2}} \tag{16a}$$

Since the nature of the curve of secondary current against mesh reactances (Fig. 12) changes completely at the point where $x_m^2 = r_{aa}r_{bb}$ this condition is defined as the condition of *critical coupling.* The condition when $x_m^2 > r_{aa}r_{bb}$ is spoken of as coupling greater than critical, whereas when $x_m^2 < r_{aa}r_{bb}$ the coupling is said to be less than critical.

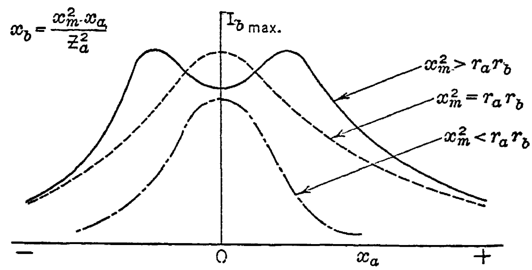


FIG. 12. Variation of I_b max with Primary Reactance for Three Degrees of Coupling

These conditions have been variously named *adequate* and *inadequate* coupling, *sufficient* and *deficient* coupling, *super* and *sub* coupling; there is no general agreement as to terminology.

When the *coupling is less than or equal to critical* there is only one inflection point on the curve of $I_{b\ \max}$ against primary reactance, which gives a maximum value for $x_a = x_b = 0$, or when each mesh is separately tuned to resonance. When the *coupling is greater than critical* the current curve shows a minimum for the condition $x_a = x_b = 0$ with maximum points on either side of zero. The proper adjustments and absolute value of currents obtainable are shown in Table 1.

Table 1. Conditions for and Values of Maximum Secondary Current in Two Mesh Circuits (For Best Possible Tuning)

	Reactance Coupling			Resistance Coupling
	$x_m^2 < r_{aa}r_{bb}$	$x_m^2 = r_{aa}r_{bb}$	$x_m^2 > r_{aa}r_{bb}$	
Optimum value for x_a.......	0	0	$\sqrt{\frac{r_{aa}}{r_{bb}}(x_m^2 - r_{aa}r_{bb})}$	0
Optimum value for x_b.......	0	0	$\sqrt{\frac{r_{bb}}{r_{aa}}(x_m^2 - r_{aa}r_{bb})}$	0
Maximum secondary current	$\frac{Ex_m}{r_{aa}r_{bb} + x_m^2}$	$\frac{E}{2\sqrt{r_{aa}r_{bb}}}$	$\frac{E}{2\sqrt{r_{aa}r_{bb}}}$	$\frac{Er_m}{r_{aa}r_{bb} - r_m^2}$
Corresponding primary current....................	$\frac{Er_{bb}}{r_{aa}r_{bb} + x_m^2}$	$\frac{E}{2r_{aa}}$	$\frac{E}{2r_{aa}}$	$\frac{Er_{bb}}{r_{aa}r_{bb} - r_m^2}$

Coupled circuits are frequently used to match two circuits with dissimilar impedances. The conditions listed in Table 1 for maximum secondary current also give the conditions for maximum power transfer and so for conjugate impedances.

TRANSMISSION-FREQUENCY CHARACTERISTIC. The most-used form of coupled circuit is that of Fig. 10, case *H*. Assuming that the distributed capacitance between windings is negligible, the transmission formula is

$$T = \frac{1}{F^2}\sqrt{\frac{L_2}{L_1}}\frac{K}{\sqrt{[d_1d_2 + K^2 - (1 - 1/F^2)^2]^2 + [(d_1 + d_2)(1 - 1/F^2)]^2}} \tag{17}$$

where $K = M/\sqrt{L_1L_2}$ is called the coefficient of coupling, and d_1 and d_2 are the decrement coefficients at resonance of the two meshes. Near resonance where $F = 1$ approximately

$$T = \sqrt{\frac{L_2}{L_1}}\frac{K}{\sqrt{2(f - f_r)/f_r]^4 + (d_1^2 + d_2^2 - 2K^2)[2(f - f_r)/f_r]^2 + (d_1d_2 + k^2)^2}} \tag{18}$$

The shape of the transmission curve as the frequency varies depends on the coefficients $(d_1^2 + d_2^2 - 2K^2)$ and $(d_1d_2 + K^2)^2$. With three independent variables resulting in two coefficients there are many solutions for a given shape. For a maximum transmission the additional condition may be specified that K is as large as possible.

SELECTIVITY. Comparing eq. (18) with eq. (4) for a single circuit it is seen that for frequencies some distance from resonance the transmission for a coupled circuit varies roughly inversely as the square of the departure from resonance, while for the single circuit the variation is roughly as the inverse first power. The transmission of two single circuits in cascade (and separated by a vacuum tube) is the product of the separate transmissions. For two circuits of decrement coefficients d_1 and d_2 this would be

$$T = \frac{1}{\sqrt{[2(f - f_r)/f_r]^4 + (d_1^2 + d_2^2)[2(f - f_r)/f_r]^2 + d_1^2 d_2^2}} \tag{19}$$

Comparison of this with eq. (18) shows that the selectivity of the coupled circuit approaches that of the cascaded single circuits as K decreases, but the transmission of the coupled circuit is decreased in the process. (For complete discussion see Purington, *Proc. I.R.E.*, Vol. 18, 983 [June 1930], and Aiken, *Proc. I.R.E.*, Vol. 25, 230 [February 1937].) Figure 13 shows the selectivity to be expected with well-designed coupled circuits. Curve C is for critical coupling; curve D for greater than critical coupling; and curves A and B for less than critical coupling.

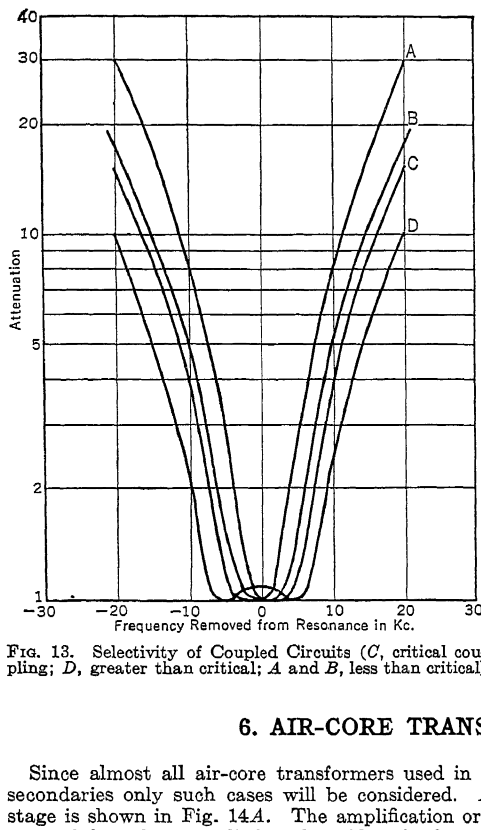


FIG. 13. Selectivity of Coupled Circuits (C, critical coupling; D, greater than critical; A and B, less than critical)

STAGGERED TUNING. The selectivity curve of a pair of coupled resonant circuits may be duplicated by the use of a two-stage amplifier having a single tuned circuit per stage, one circuit being tuned above and the other below the desired mean frequency. Similarly any desired number of single-tuned-circuit stages may be used to obtain an overall selectivity curve with a flat top and steep sides. The tuning points are distributed systematically across the pass bands. The effective Q is highest on the circuits tuned near the cutoff frequency. See Section 7, article 13; also the literature (particularly V. D. Landon, Cascade Amplifiers with Maximal Flatness, *RCA Rev.*, Vol. V, No. 3–4 [January–April 1941]).

6. AIR-CORE TRANSFORMERS

Since almost all air-core transformers used in communication circuits employ tuned secondaries only such cases will be considered. A simple circuit for a tuned amplifier stage is shown in Fig. 14*A*. The amplification or gain of such a stage is defined as the ratio of the voltage applied to the grid of the first tube to the voltage delivered to the grid of the second tube. The value of the gain at resonance is

$$\frac{E_2}{E_1} = \frac{\mu\omega^2 M^2/r_s}{(\omega^2 M^2/r_s) + r_p} \times \frac{L_s}{M} \tag{20}$$

where μ is the amplification factor of the tube.

In this equation $\omega^2 M^2/r_s$ represents the impedance reflected into the primary circuit by the tuned secondary circuit. It is assumed in this equation that the primary reactance is negligible.

This equation may be written in a simpler algebraic form but is most easily remembered and used in the form shown. Keeping in mind that $\omega^2 M^2/r_s$ is the load on the tube it is evident that $\mu E_1 \dfrac{\omega^2 M^2/r_s}{(\omega^2 M^2/r_s) + r_p}$ is the voltage drop across the plate load. Multiplying this voltage by the ratio of transformation L_s/M gives E_2, the output voltage.

As the mutual reactance between the primary and secondary circuits is varied, a maximum of gain is obtained when the value of ω^2M^2/r_s is equal to r_p. This constitutes matching the impedance of the tube.

In modern tubes of the screen-grid or of the pentode type, the plate impedance (r_p) is of such a high magnitude that it is ordinarily impracticable to match its impedance. In fact, ordinarily the load in the plate circuit is so small as to have a negligible effect upon the plate current, in which case the formula for gain becomes $\dfrac{E_2}{E_1} = \dfrac{s_m\omega^2M^2}{r_s}\dfrac{L_s}{M}$, where s_m is the mutual conductance, or transconductance, of the tube. Expressing this equation in words, the gain of a tuned amplifier, employing a tube of very high plate impedance, is equal to the transconductance, multiplied by the load impedance, and by the transformation ratio.

The equation may be written in another form, $\dfrac{E_2}{E_1} = s_m \dfrac{\omega L}{r_s} \omega M$. That is, the gain is equal to the transconductance, divided by the power factor of the secondary, and multiplied by the mutual reactance.

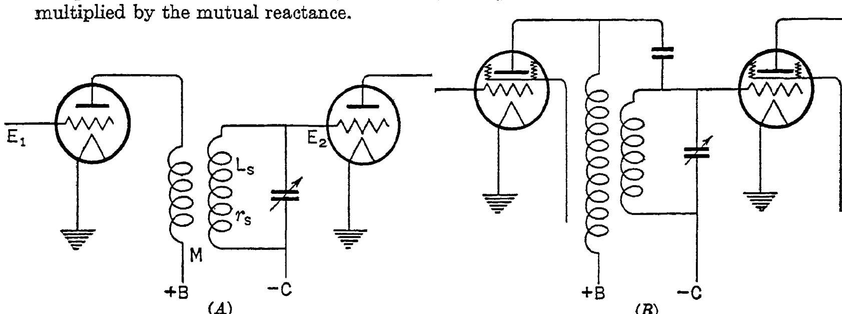


FIG. 14. Tuned Amplifier Circuits

VARIATION OF GAIN WITH FREQUENCY OF RESONANCE. The gain of an amplifier of this type varies considerably with frequency as it is tuned over the frequency band. The frequency squared occurs in the numerator, causing the gain to tend to be larger at the high-frequency end of the tuning range. This tendency is partly counterbalanced by the normal variation of the circuit resistance with frequency. The rate of change of resistance varies considerably from coil to coil, but the resistance varies faster than the first power of the frequency, and slower than the second power. If the resistance varies directly with the frequency the gain is proportional to the frequency of resonance. However, if the resistance is proportional to the square of the frequency the gain is constant. Usually the resistance rises only slightly faster than the frequency, causing the gain to rise with frequency.

EFFECT OF LEAKAGE REACTANCE. With tight coupling it is justifiable to neglect the primary reactance and the tube's plate-filament capacitance, because the value of the reflected impedance exceeds the primary reactance so greatly.

However, when the coupling is only moderate, the plate-filament capacitance and the primary reactance are no longer negligible. The effect is to increase the gain, above that of the formula, especially at high frequencies.

In practice, the transformer is frequently of such a design that the primary reactance is far from negligible. The gain formula then becomes somewhat too complicated for practical use. Cut-and-try methods are the rule for designing such transformers.

TUNED R-F TRANSFORMER EMPLOYING COMPOUND COUPLING. A common design of transformer in broadcast receivers employing screen-grid tubes is that shown in Fig. 14*B*. It employs a primary of such a high inductance that its primary is resonant in conjunction with the output capacitance of the tube, at a frequency below the broadcast frequency spectrum, 450 kc and below. The secondary coil is of the usual type, but the per cent coupling is quite low, usually between 15 and 25 per cent. Sufficient capacitance coupling is used to give the desired high-frequency gain.

The performance of this transformer is an improvement over that of the low-inductance primary type, in that the gain may be made more nearly constant throughout the tuning range.

The average voltage gain in this type of transformer is somewhat less, but the great amplifying ability of screen-grid tubes makes it unnecessary to obtain the greatest possible gain in transformers.

7. THREE-WINDING TRANSFORMERS (HYBRID COILS)

It is frequently desirable in electric circuits that currents in one portion of a circuit shall induce voltage in all branches of the circuit except certain designated ones, in which no voltage is to be introduced. This may be accomplished by means of an *impedance bridge* (Fig. 15) which consists of six adjustable impedances arranged in the form of a Wheatstone bridge. If the impedances are so adjusted that for a certain frequency $z_A/z_B = z_C/z_D$ a voltage of that frequency introduced at E will cause no current in the arm G, and vice versa.

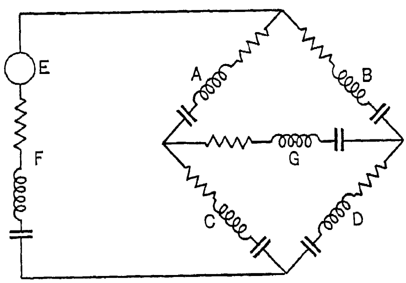


FIG. 15. Impedance Bridge Circuit

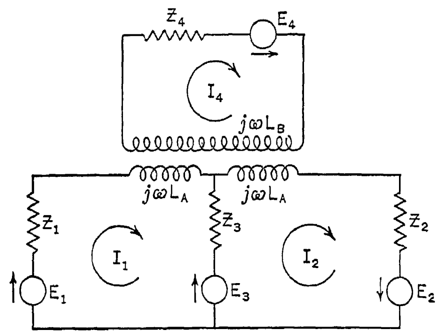


FIG. 16. Schematic Circuit of Three-winding Transformer with Load

A more widely used means of blocking voltages out of a particular branch is a combination of three coils such as that shown in Fig. 16; it is optionally called a *three-winding transformer* or a *hybrid coil.* If the two coils L_a are wound series aiding and the transformer is well made so that the winding resistances may be neglected and the coefficients of coupling are practically unity, then a voltage E_1 will cause no current through E_2 (and vice versa) provided $z_3 = z_4/N^2$, where N is the turn ratio between one of the L_a coils and the L_b coil. Also if $z_1 = z_2$ a voltage E_3 will cause no current through E_4, and vice versa. If all these conditions are fulfilled, then power from E_1 or E_2 will divide equally between z_3 and z_4 and power from E_3 and E_4 will divide equally between z_1 and z_2. Such transformers are extensively used in bidirectional amplifiers (see p. 7-13); also this circuit is the basic circuit for all neutralizing circuits. (See p. 7-29)

BIBLIOGRAPHY

Adams, J. J., Undercoupling in Tuned Coupled Circuits to Realize Optimum Gain and Selectivity, *Proc. I.R.E.*, Vol. 29, 277 (May 1941).
Aiken, C. B., Two-mesh Tuned Coupled Circuit Filters, *Proc. I.R.E.*, Vol. 25, 230 (February 1937).
Blanchard, J., The History of Electrical Resonance, *B.S.T.J.*, Vol. 20, 415 (October 1941).
Everitt, W. L., *Communication Engineering.* McGraw-Hill, New York (1937).
Johnson, K. S., *Transmission Circuits for Telephonic Communication.* Van Nostrand, New York (1925).
King, R., The Application of Low-frequency Circuit Analysis to the Problem of Distributed Coupling in Ultra-high-frequency Circuits, *Proc. I.R.E.*, Vol. 27, 715 (November 1939).
King, R., A Generalized Coupling Theorem for Ultra-high-frequency Circuits, *Proc. I.R.E.*, Vol. 28, 84 (February 1940).
Korman, N. I., Coupled Resonant Circuits for Transmitters, *Proc. I.R.E.*, Vol. 31, 28 (January 1943).
Landon, V. D., Cascade Amplifiers with Maximal Flatness, *R.C.A. Review*, Vol. 5, No. 3–4 (January 1941).
Lee, R., A Practical Analysis of Parallel Resonance, *Proc. I.R.E.*, Vol. 21, 271 (February 1933).
McElroy, P. K., Designing Resistive Attenuating Networks, *Proc. I.R.E.*, Vol. 23, 213 (March 1935).
McIlwain, K., and J. G. Brainerd, *High-frequency Alternating Currents.* John Wiley, New York (1939).
Mead, S. P., Phase Distortion and Phase Distortion Correction, *B.S.T.J.*, Vol. 7, 195 (April 1928).
Nelson, J. R., A Theoretical Comparison of Coupled Amplifiers with Staggered Circuits, *Proc. I.R.E.*, Vol. 20, 1203 (July 1932).
Purington, E. S., Single- and Coupled-circuit Systems, *Proc. I.R.E.*, Vol. 18, 983 (June 1930).
Sherman, J. B., Some Aspects of Coupled and Resonant Circuits, *Proc. I.R.E.*, Vol. 30, 505 (November 1942).
Terman, F. E., *Radio Engineering.* McGraw-Hill, New York (1937).
Zobel, O. J., Distortion Correction in Electrical Circuits, *B.S.T.J.*, Vol. 7, 438 (July 1928).

Also many articles in *Proc. I.R.E.*, *J. A.I.E.E.*, and *Experimental Wireless and Wireless Eng.*

TRANSFORMERS WITH IRON CORES

By A. J. Rohner

8. AUDIO-FREQUENCY TRANSFORMERS

The Function of the audio-frequency transformer is to couple various circuits, at audio frequencies, over a considerable range of the audio-frequency band. It may be used as an impedance-matching device, as a means of isolating circuits, or as a means of obtaining phase reversal. When it couples a voltage source, such as a microphone, a phonograph pick-up, or a telephone line, to the grid of a vacuum tube, it is usually called an **input transformer.** If it couples the plate of one tube to the grid of another, it is called an **interstage transformer.** If it couples the plate of a vacuum tube to some sort of load, such as a loudspeaker, indicating meter, or telephone line, it is referred to as an **output transformer.** The **modulation transformer** is a special case of the output transformer, in which the load is the plate of a radio-frequency amplifier tube. If the transformer is used to match telephone lines of unequal impedance, or to isolate lines of equal impedance, it is called a **line transformer.** There are, of course, many variations of the above-mentioned applications.

The audio-frequency transformer is constructed much like a power transformer, but there are two distinct points of difference. Power transformers are usually one-frequency devices, whereas audio transformers must operate over a wide band of frequencies. In high-fidelity systems, for example, a range of from 30 to 15,000 cycles may be required. In systems where intelligibility of speech is all that is necessary, 300 to 3000 cycles may suffice. Then, too, the power transformer works from a voltage source having good regulation; that is, the impedance of the source, or "generator impedance," is negligible as compared with the load impedance. An audio transformer always works from a voltage source having poor regulation, the generator impedance often being equal to the load impedance, and in pentode or Class B power amplifiers, being greater than the load impedance.

These two factors, wide frequency range, and high generator impedance, place severe restrictions on the constants of an audio transformer. Primary inductance must be high; leakage inductance and distributed and other capacitances must be low. See G. Koehler, Design of Transformers for Audio-frequency Amplifiers with Preassigned Characteristics, *Proc. I.R.E.* (December 1928); P. W. Willans, Low-frequency Intervalve Transformers, *I.E.E. J.* (October 1926).

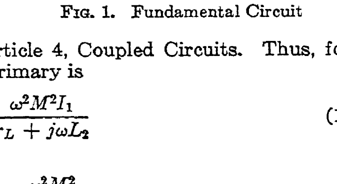


FIG. 1. Fundamental Circuit

COUPLED CIRCUITS. The design of an audio-frequency transformer requires the solution of coupled circuits, of which the transformer is a part, and including the generator and the load impedances. This solution may be carried out by the classical method, using mutual inductance. See article 4, Coupled Circuits. Thus, for the circuit of Fig. 1, the voltage impressed on the primary is

$$E_1 = r_pI_1 + j\omega L_1I_1 + \frac{\omega^2M^2I_1}{r_L + j\omega L_2} \tag{1}$$

The impedance "looking in" to the primary is

$$Z_1 = \frac{E_1}{I_1} = r_p + j\omega L_1 + \frac{\omega^2M^2}{r_L + j\omega L_2} \tag{2}$$

The voltage across the load, r_L, is

$$E_2 = r_LI_2 = \frac{\omega MI_1r_L}{r_L + j\omega L_2} \tag{3}$$

This method becomes quite complicated, however, when all the factors affecting audio transformers are considered: core loss, winding resistances, distributed capacitance, interwinding capacitance, turns ratio, etc. The first step in simplifying the analysis of the audio transformer and its associated circuits is to convert them to an equivalent direct-connected network. See J. H. Morecroft, *Principles of Radio Communication*, 2nd Ed., pp. 95–105.

EQUIVALENT DIRECT-CONNECTED NETWORK. Considering the transformer shown in Fig. 1 as a unity-ratio transformer, $L_1 = L_2$. Then $kL_1 = kL_2$, where k = coefficient of coupling, and $M = k\sqrt{L_1L_2} = kL_1$, where M is the mutual inductance.

The leakage inductance of the primary, or that portion of the primary which is not coupled to the secondary, equals $(1 - k)L_1$. Similarly, the portion of the secondary not coupled to the primary equals $(1 - k)L_2$, which is, of course, equal to $(1 - k)L_1$. The circuit of Fig. 1 can now be replaced by a direct-connected network as shown by Fig. 2. The common impedance is kL_1.

It may be shown that this circuit is the exact equivalent of Fig. 1. For example, the input impedance equals Z_1 and

$$Z_1 = r_p + j\omega L_1 + \frac{\omega^2 k^2 L_1^2}{r_L + j\omega L_1}$$

Substituting M for kL_1, and L_2 for L_1

$$Z_1 = r_p + j\omega L_1 + \frac{\omega^2 M^2}{r_L + j\omega L_2}$$

which is the expression derived for the circuit of Fig. 1.

The equivalent network of an inequality ratio transformer may be referred to either the primary or secondary circuit, provided the correct transformations are made. To reflect the secondary constants to the primary, for example, it is necessary to multiply all the impedances on the secondary side by the ratio L_1/L_2. With iron-core transformers, the ratio L_1/L_2 is for all practical purposes equal to the square of the turns ratio. Voltages are transformed by the turns ratio, currents by the inverse of the turns ratio.

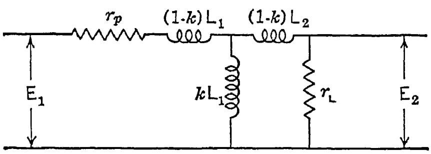


Fig. 2. Equivalent Network

If N_p = primary turns, and N_s = secondary turns, the secondary constants, referred to the primary, are

$$L_2' = \left(\frac{N_p}{N_s}\right)^2 L_2 \qquad r_2' = \left(\frac{N_p}{N_s}\right)^2 r_2 \qquad C_2' = \left(\frac{N_s}{N_p}\right)^2 C_2$$

$$E_2' = \left(\frac{N_p}{N_s}\right) E_2 \qquad I_2' = \left(\frac{N_s}{N_p}\right) I_2 \qquad r_L' = \left(\frac{N_p}{N_s}\right)^2 r_L$$

After these conversions are made, the transformer becomes a unity-ratio transformer and can be replaced by an equivalent direct-connected network.

COMPLETE EQUIVALENT NETWORK. Figure 3 shows the complete equivalent network of an audio-frequency transformer, referred to the primary, in which E_1 is the generator voltage (μEg in the case of a vacuum tube); r_p is the generator resistance; C_1 is the distributed capacitance of the primary winding, plus any additional capacitance across that winding; r_1 is the primary winding resistance; r_c is the core-loss resistance; r_2' is the secondary winding resistance, referred to the primary; C_2' is the distributed capacitance of the secondary winding, plus additional capacitances across that winding within the transformer itself, referred to the primary; and r_L', C_L', and E_2' are the load resistance, load capacitance, and load voltage, all referred to the primary.

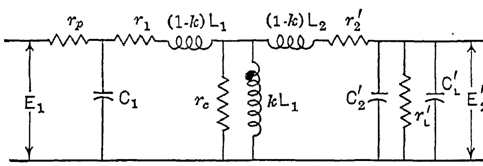


Fig. 3. Equivalent Network of Audio Transformer

The "additional capacitances" spoken of may be capacitances of the windings to ground, or capacitance between windings. For example, a single-ended interstage transformer might have its primary on the inside, next to the core, and secondary on the outside, wound over the primary. If the primary finish is connected to $+B$, and the secondary start is connected to ground, there is no a-c potential between the adjacent surfaces of the two windings, so that capacitance between the two windings has no effect. However, the primary start would go to plate, and the start layer of the winding has capacitance to the core. This capacitance is an additional capacitance across the primary winding. Similarly, the secondary finish would go to grid, and any capacitance existing between the finish layer of the secondary and the core or case would be an additional capacitance across the secondary winding. If the connections to both windings were reversed, making primary start $+B$, and secondary finish ground, those capacitances would have no effect. The capacitance between windings, however, would be of great importance. This would be somewhat equivalent to an additional capacitance across the winding having the most turns. Use of a grounded electrostatic shield between windings will largely eliminate capacitance between windings, but, of course, it adds additional capacitances to ground.

Solution of the circuit of Fig. 3 will give the characteristics of the audio transformer, such as (1) voltage ratio as a function of frequency; (2) primary impedance as a function of frequency; (3) phase shift as a function of frequency; (4) efficiency.

SIMPLIFIED NETWORK AT LOW FREQUENCIES. The network of Fig. 3 can be greatly simplified for practical design purposes by considering, first, the effect of frequency upon the relative importance of the various constants, and second, the particular application in which the transformer is used.

At low frequencies, leakage inductance and all the capacitances can be ignored. Furthermore the coefficient of coupling, k, of most audio transformers is above 0.995, so that kL_1 may be taken as L_1 with an error of less than 1 per cent. The low-frequency characteristics of any audio transformer are determined by the primary inductance and the various resistances of the network, as shown in Fig. 4.

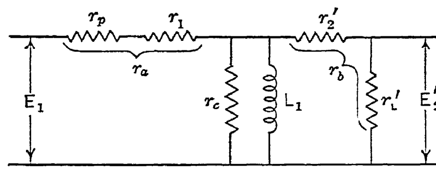


FIG. 4. Network at Low Frequencies

But Fig. 4 can be simplified still further. The resistances can be lumped together into a single resistor, as shown in Figs. 5 and 6. Let $r_a = r_p + r_1$ and let $r_b = r_2' + r_L'$. Then the lumped resistance, R, is equivalent to r_a, r_b, and r_c in parallel, or

$$R = \frac{r_a r_b r_c}{r_a r_b + r_a r_c + r_b r_c} \tag{4}$$

If the input voltage, E_1, is multiplied by the attenuation of the resistance network, the circuit of Fig. 5 becomes identical, in its output voltage and phase shift, with Fig. 4. The new value for the input voltage, of Fig. 5, is

$$E = \frac{r_L' r_c E_1}{r_a r_b + r_a r_c + r_b r_c} \tag{5}$$

Figure 6 is simply a redrawing of Fig. 5, showing the single resistor and the new input voltage. Figure 6 is an exact equivalent of Fig. 4 as far as voltage output and phase shift are concerned. The voltage output and the phase shift, at low frequencies, of *all* audio transformers are given by this simple circuit. Figure 7 shows those characteristics, as a function of $2\pi f L_1/R$.

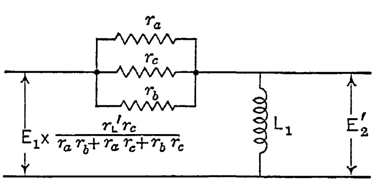


FIG. 5. Method of Combining Resistances

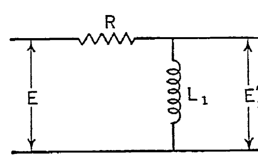


FIG. 6. Circuit of Audio Transformer at Low Frequencies

At some frequency, $2\pi f L_1 = R$, so that $2\pi f L_1/R = 1$. Let this particular frequency be called f_1. This is the low frequency at which the response has dropped 3 db from its value in the middle-frequency range. At twice this frequency, $2\pi f L_1/R = 2$. At three times this frequency, $2\pi f L_1/R = 3$, etc. At any frequency, f, $2\pi f L_1/R = f/f_1$. The curves given in Fig. 7, as a function of $2\pi f L_1/R$, are at the same time frequency characteristics, the frequency being expressed in terms of the reference frequency, f_1.

In order to determine the proper value of primary inductance, it is necessary to know what drop in secondary voltage is permissible, at some specified low frequency, as compared with the voltage in the middle-frequency range. For example, 1-db drop at 100 cycles might be given as the requirement of low-frequency response. From Fig. 7, the ratio of $2\pi f L_1/R$ is found, which gives this particular drop. As a first approximation, the winding and core-loss resistances may be neglected, so

$$R \cong \frac{r_p r_L'}{r_p + rL'} \tag{6}$$

Then

$$L_1 = \frac{2\pi f L_1}{R} \times \frac{R}{2\pi f} \tag{7}$$

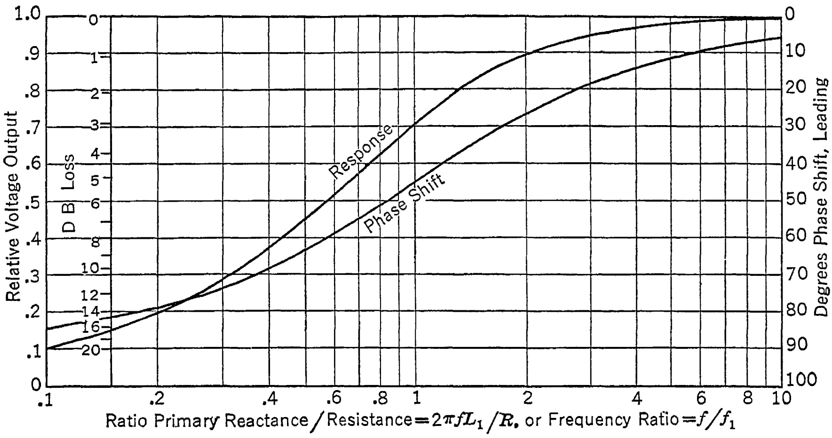


Fig. 7. Response and Phase Shift at Low Frequencies

SIMPLIFIED NETWORK AT MIDDLE FREQUENCIES. In this frequency range, all reactance elements become negligible, and the transformer reduces to a network of resistances, as shown in Fig. 8. In this range, phase shift is practically zero. This is the "flat" portion of the frequency-response characteristic, the secondary voltage being

$$E_2 = \frac{N_s}{N_p} \times E_1 \times \frac{r_L' r_c}{r_a r_b + r_a r_c + r_b r_c} \tag{8}$$

and the efficiency of the transformer being, to a very close approximation,

$$\text{Efficiency} = \frac{r_L'}{r_1 + r_2' + r_L' + (r_L'/r_c)(2r_1 + 2r_2' + r_L')} \tag{9}$$

SIMPLIFIED NETWORK AT HIGH FREQUENCIES. The shunting effect of the primary inductance is negligible at high frequencies. See Fig. 7. If the generator resistance, r_p, or the reflected load resistance, r_L', is less than 20.000 ohms, the primary capacitance, C_1, may be neglected. Most audio transformers fall in this class. Similarly, considering the secondary side of the transformer, if the *reflected* generator resistance, or the load resistance, is less than 20,000 ohms, the secondary capacitance, C_2, may be neglected. Though this is usually true of output transformers, it is seldom true of input or interstage transformers.

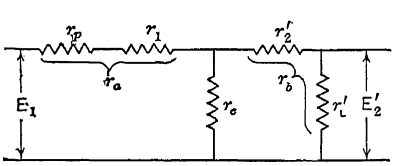


Fig. 8. Equivalent Network at Middle Frequencies

The core-loss resistance, r_c, has little effect at high frequencies beyond reducing the secondary voltage by a few per cent. The per cent voltage drop caused by core loss, using the symbols of Fig. 8, is

$$\text{Core loss drop} = \frac{100}{1 + r_c/r_b + r_c/r_a} \text{ per cent} \tag{10}$$

This usually amounts to 2 or 3 per cent. As far as the shape of the response curve, or the amount of phase shift, is concerned, core loss may be neglected.

Neglecting primary inductance, primary capacitance, and core loss, the equivalent circuit, at high frequencies, becomes as shown in Fig. 9. The term $2(1 - k)L_1$, is called the leakage inductance referred to the primary and is usually designated by L_e.

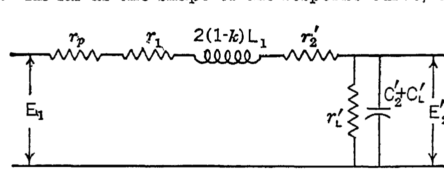


Fig. 9. Equivalent Network at High Frequencies

9. OUTPUT TRANSFORMERS

The function of the output transformer is to transfer power from the plate, or plates, of vacuum tubes to a load, such as a loudspeaker, an indicating meter, or a line. It provides the necessary impedance transformation, and it isolates the load from the d-c potential and current of the plate circuit. Efficiency is usually important, and the transformer must meet a prescribed frequency-response characteristic.

TURNS RATIO is determined by the plate load recommended for the tube, or tubes, by the tube manufacturer, r_L', and the actual load resistance, r_L.

$$\frac{N_p}{N_s} = \sqrt{\frac{r_L'}{r_L}} \tag{11}$$

FREQUENCY RESPONSE is controlled by the amount of primary inductance, at low frequencies, and by the amount of leakage inductance, at high frequencies. The allowable drop at low frequencies fixes a minimum value of primary inductance. See Fig. 7 and eqs. (6) and (7).

The equivalent network of the output transformer at high frequencies is given by Fig. 10, which is the same as Fig. 9 except that all capacitances have been omitted. The

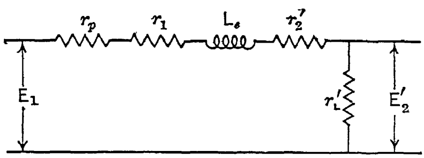


FIG. 10. Output Transformer at High Frequencies

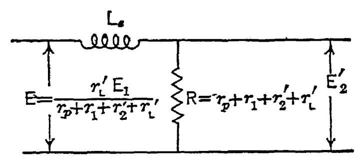


FIG. 11. Circuit Equivalent to Fig. 10

capacitance of an ordinary audio transformer winding will be about 100 $\mu\mu f$, more or less, depending upon the coil construction. At 10,000 cycles, this is 160,000 ohms of capacitive reactance. The load on an output transformer is seldom over a few thousand ohms and may be as low as 3 or 4 ohms. The shunting effect of the secondary capacitance is therefore negligible, even at the highest audio frequencies. The reflected load, looking into the primary, is seldom greater than 20,000 ohms, e.g., pushpull 6F6 tubes require 10,000 ohms. Again, primary capacitance is negligible.

The circuit of Fig. 11 is the exact equivalent of that of Fig. 10, as far as voltage output and phase shift are concerned. The voltage output and the phase shift of *all* output

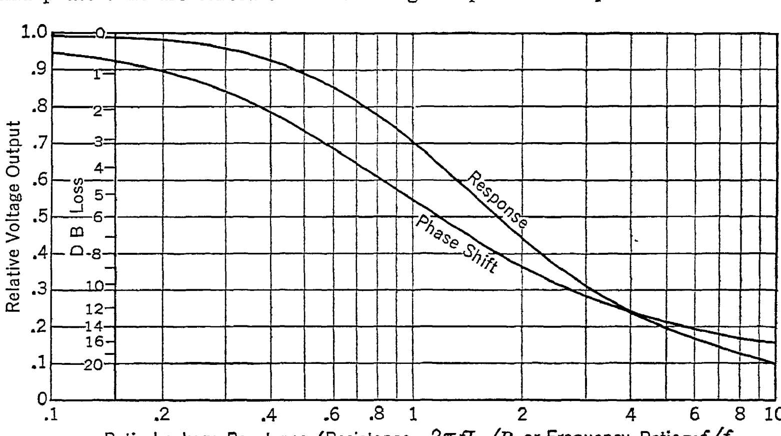


FIG. 12. Response and Phase Shift. Output Transformers at High Frequencies

transformers at high frequencies are given by this simple circuit, provided that the load is resistive and the load and reflected load are less than 20,000 ohms. Figure 12 shows those characteristics as a function of $2\pi fL_e/R$.

At some frequency, $2\pi f L_e = R$, so that $2\pi f L_e/R = 1$. Call this particular frequency f_2. This is the high frequency at which the response has dropped 3 db from its value in the middle-frequency range. At twice this frequency $2\pi f L_e/R = 2$. At three times this frequency, $2\pi f L_e/R = 3$, etc. At any frequency, f, $2\pi f L_e/R = f/f_2$. The curves given in Fig. 12, as a function of $2\pi f L_e/R$, are at the same time frequency characteristics, the frequency being expressed in terms of the reference frequency, f_2.

To determine the allowable amount of leakage inductance, it is necessary to know what drop in secondary voltage is permissible, at some specified high frequency, as compared with the voltage in the middle-frequency range. From Fig. 12, the ratio of $2\pi f L_e/R$ which gives this drop is found.

$$R = r_p + r_1 + r_2' + r_L'$$

Then

$$L_e = \frac{2\pi f L_e}{R} \times \frac{R}{2\pi f} \tag{12}$$

See F. E. Terman, *Radio Engineering*, 2nd Ed., pp. 293–299; L. A. Kelley, Transformer Design, *Rad. Engrg.* (December 1934, February 1935); F. E. Terman and R. E. Ingebretsen, Output Transformer Response, *Electronics*, January 1936; and *Magnetic Circuits and Transformers*, staff of M.I.T., pp. 472–486.

EFFICIENCY of output transformers is usually between 80 and 90 per cent, though it may be as low as 60 per cent for cheap or poorly designed transformers. Maximum efficiency is obtained, for a given physical size and core material, when copper loss = core loss, and when primary copper loss = secondary copper loss. Such a balance of losses is not always possible, however. If the secondary resistance, r_2, is made 5 per cent of the load resistance, r_L; if the primary resistance, r_1, is made 5 per cent of the reflected load resistance, r_L'; and if the core-loss resistance, r_c, is made 10 times the reflected load, r_L', the losses will be very nearly balanced, and the efficiency will be 82 per cent. See eq. (9).

LOUDSPEAKER LOAD. The analysis of output transformers given above has assumed a constant, resistive load. If the load impedance is not constant throughout the frequency range, the frequency-response characteristic will not be flat but will rise and fall where the load impedance rises and falls. This effect is not very pronounced with Class A triode amplifiers. However, with pentode or with Class B triode amplifiers, the output voltage is approximately proportional to the load impedance. When such tubes are used to drive the conventional moving-coil loudspeaker, a flat frequency-response characteristic is no longer obtained.

The impedance-frequency characteristic of a moving-coil loudspeaker is characterized by a low-frequency resonance peak and by a rise in impedance with frequency at the higher audio frequencies. In the neighborhood of 400 cycles, the loudspeaker impedance is the minimum and is resistive. This minimum, resistive impedance should be used as the basis of calculating the turns ratio and the efficiency of the transformer.

Flattening of the frequency-response curve by mismatching, that is, by using a value for r_L perhaps twice the value of the actual minimum, resistive impedance, in order to favor the low and high frequencies at the expense of the middle frequencies, is somewhat effective when used with Class A triode amplifiers. It does not level the voltage characteristic but does tend to level the power output over a wider frequency range. Mismatching is futile, however, when used with pentode or Class B amplifiers.

The rise in impedance at the high frequencies may be offset by connecting a capacitor, or a resistor and capacitor in series, across the primary of the output transformer. The low-frequency peak is best controlled by acoustical damping of the loudspeaker itself.

PUSHPULL OUTPUT TRANSFORMER, CLASS A. No special problems are introduced by pushpull operation of the output transformer if the amplifier is Class A. (See Section 7, Amplifiers.) The generator resistance, r_p, is twice the plate resistance of one tube. The reflected load, r_L', is the recommended tube load, plate-to-plate. In fact, the design of the transformer is simpler, for pushpull operation, because the d-c plate currents of the two tubes flow in opposite directions in the transformer windings, so that d-c magnetization of the core is canceled out. This results in higher primary inductance and a better low-frequency response.

Data are furnished by the tube manufacturers on optimum plate-to-plate load. The method of arriving at this optimum load is discussed by B. J. Thompson, Graphical Determination of Performance of Pushpull Audio Amplifiers, *Proc. I.R.E.*, April 1933.

PUSHPULL OUTPUT TRANSFORMER, CLASS B. Class B operation imposes special requirements on the output transformer, because one half of the primary works during one half-cycle, and the other during the other half-cycle. It is important that the two halves of the primary be closely coupled, so that the cross-over from one to the other may be accomplished smoothly and without introducing transients. It is also important that each half of the primary be coupled equally to the secondary. Otherwise the high-frequency response of the transformer will not be the same for both half-cycles, which will produce even harmonics in the output wave. In general, leakage inductance should be kept to the minimum between the windings of a Class B output transformer, even beyond the requirements of frequency response.

These requirements of low leakage and equal coupling are met by using the coil construction illustrated in Fig. 13.

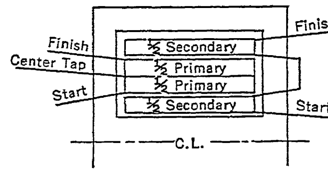


Fig. 13. Arrangement of Windings, Class B Output Transformer

THE MODULATION TRANSFORMER is an output transformer, which has as its load the plate of a Class C radio-frequency amplifier. This load is resistive and is usually of the order of a few thousand ohms. (See Section 7, Modulators.) The audio-frequency generator is usually a Class B amplifier, so that the discussion of Class B output transformers given above applies to modulation transformers.

The secondary often is required to carry the d-c current of the Class C amplifier. This produces a d-c magnetization of the core, which must be considered when designing the transformer, because of its effect upon inductance and low-frequency response as well as upon heating. Core saturation, due to d-c and a-c magnetization, is usually an important factor in modulation transformers. It is most serious at the lowest frequency of the frequency range, since there the a-c flux density is greatest. The effects of d-c magnetization of the core upon inductance and upon saturation are discussed in Section 3, Ferrous-cored Inductors.

Modulation transformers often work at high power levels, of the order of hundreds or thousands of watts. Heating is an important consideration, as with power transformers. Usually, too, high voltages are applied to the modulation transformer. The primary center tap and one end of the secondary winding are connected to the d-c plate supply and must be insulated to withstand its voltage. The ends of the primary and the other end of the secondary, all of which are connected to plates, must withstand twice the d-c plate supply voltage, since they have audio-frequency voltage, additional to the d-c voltage, and of a peak value approximately equal to the d-c voltage.

Since the primary center tap and one end of the secondary are often connected to a common point, the *B* supply, and since the load resistance of the Class C amplifier is of the same order of magnitude as the reflected load on one tube of the Class B amplifier, there is a strong temptation to make the modulation transformer an auto transformer, having one half of the primary common with the secondary. The saving in size, cost, and insulation by such construction is very great. Invariably, however, such construction leads to trouble due to the unequal coupling between the two halves of the primary and the secondary. (See above, Pushpull, Class B.)

For a practical analysis of Class B modulation, and driver, transformers, see J. Kunz, Transformers for Class B Modulators, *Radio Engrg.*, July 1934.

THE LINE TRANSFORMER is not an output transformer, as it does not work out of the plate of a vacuum tube. However, the analysis presented above for the output transformer applies equally well to the line transformer. Turns ratio is determined in the same manner. Frequency response and phase shift are governed by the same factors, viz., primary and leakage inductances.

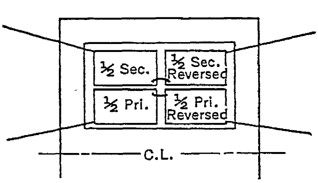


Fig. 14. Balanced Coil Construction

Line transformers usually operate at low power levels, of the order of 6 milliwatts, or less, so that shielding from stray magnetic fields may be necessary.

Often, too, it is necessary to balance the line to ground. This means that the capacitance from one end of the primary to ground shall equal that from the other, and the capacitance from one end of the secondary to ground shall equal that from the other. A symmetrical coil construction, as shown in Fig. 14, accomplishes this purpose.

10. INPUT AND INTERSTAGE TRANSFORMERS

The function of the input transformer is to couple an audio-frequency voltage source, such as a microphone, phonograph pick-up, or telephone line, to the grid of a vacuum tube. That of the interstage transformer is to couple the plate of one tube to the grid of another. Either type must conform to a predetermined frequency-response characteristic and must furnish the greatest possible voltage amplification consistent therewith.

With either type of transformer, the load consists of the grid circuit of a vacuum tube (or tubes), and its impedance is very high, frequently of the order of megohms. It is often no more than the input capacitance of the tube, although sometimes a resistor of 100,000 to 500,000 ohms is placed across the secondary, also. With such a high-impedance load, the secondary winding capacitance is not negligible. The secondary winding capacitance and the load capacitance, together, largely control the high-frequency response and the turns ratio of the transformer.

FREQUENCY CHARACTERISTICS, LOW AND MIDDLE FREQUENCIES. The same analysis applies to input and interstage transformers, at these frequencies, as applies to other audio transformers. See above, Simplified Network at Low Frequencies and Middle Frequencies. If there is no resistance load on the secondary, that is, if $r_L' = \infty$, the equations given become simpler. Thus, eq. (8), for the secondary voltage, becomes

$$E_2 = \frac{N_s}{N_p} \times E_1 \times \frac{r_c}{r_a + r_c} \tag{13}$$

and eq. (9), for efficiency, becomes

$$\text{Efficiency} = 0 \tag{14}$$

FREQUENCY CHARACTERISTIC, HIGH FREQUENCIES. Figure 9 shows the equivalent network at high frequencies. Considering first the case when there is no secondary resistance load, the equivalent circuit becomes that shown in Fig. 15. This is a simple series resonant circuit. It is convenient to express the performance of this circuit by means of a family of curves, plotting voltage ratio vs. frequency, as shown by Fig. 16. It is seen that the shape of the desired frequency-response characteristic determines the value of a constant, N, while the position of the desired characteristic on the frequency band determines the value of the resonant frequency, f_0. Then

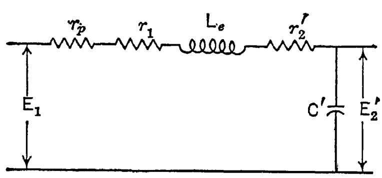


FIG. 15. Input or Interstage Transformer at High Frequencies

$$\sqrt{\frac{L_e}{C'}} = N(r_p + r_1 + r_2') \tag{15}$$

and

$$\sqrt{L_e C'} = \frac{1}{2\pi f_0} \tag{16}$$

From these equations, the value of the leakage inductance, L_e, and of the reflected secondary capacitances, C', which will give the desired frequency response, can be calculated. If the amplification at high frequencies is to be substantially constant, the value of the constant N must lie between 0.75 and 1.0. See F. E. Terman, *Radio Engineering*, 2nd Ed., pp. 188–202; *Gen. Elec. Tech. Report* 19366, Audio Transformer Design (July 1930); and *Magnetic Circuits and Transformers*, staff of M.I.T., pp. 486–494.

Considering the case in which a resistance load is placed across the secondary, in addition to the tube input and secondary winding capacitances, the equivalent network is that of Fig. 9. The resistance load will reduce the leakage resonance peak, at the same time lowering the voltage output in the middle-frequency range in accordance with eq. (8). However, the effect is more pronounced in the region of leakage resonance, so that an overall flattening of the response characteristic results. See Terman, p. 199.

No single family of curves can be drawn which will show the performance of this circuit. However, two methods of attacking the problem have been described. One method uses a master chart showing the response at the leakage-resonance frequency. This chart enables a designer to pick out a value of secondary loading resistance that will give the desired response at resonance. He then computes the response at a few other points, sufficient for plotting the frequency-response characteristic. See P. W. Klipsch, A.F. Amplifier Circuits Using Transformers, *Proc. I.R.E.*, February 1936.

Another method of solving the circuit of Fig. 9 is to draw several families of curves, each family representing some fixed relationship between the various constants. This method, with six such families of curves, is described by J. G. Story, Design of A.F. Input and Intervalve Transformers, *Wireless Engr.*, February 1938.

THE TURNS RATIO, of an input or interstage transformer is determined by the secondary capacitance, that is, by the sum of the secondary winding capacitance, C_2, and the input capacitance of the tube, C_L. Call this total capacitance C, and let its value reflected to the primary be C'. The correct value of C' is found from the frequency-response requirements, as described above. Then

$$\frac{N_s}{N_p} = \sqrt{\frac{C'}{C}} \tag{17}$$

The capacitance, distributed, and to ground, of the secondary winding cannot be calculated accurately until the design of the transformer has been completed. As a first approximation, for finding the turns ratio, a value of 50 $\mu\mu f$ may be used. The input capacitance of the tube is given by the formula

$$C_L = C_{gf} + C_{gp}(1 + \mu)$$

where C_{gf} = static capacitance between grid and filament, C_{gp} = static capacitance between grid and plate, and μ = effective amplification of tube.

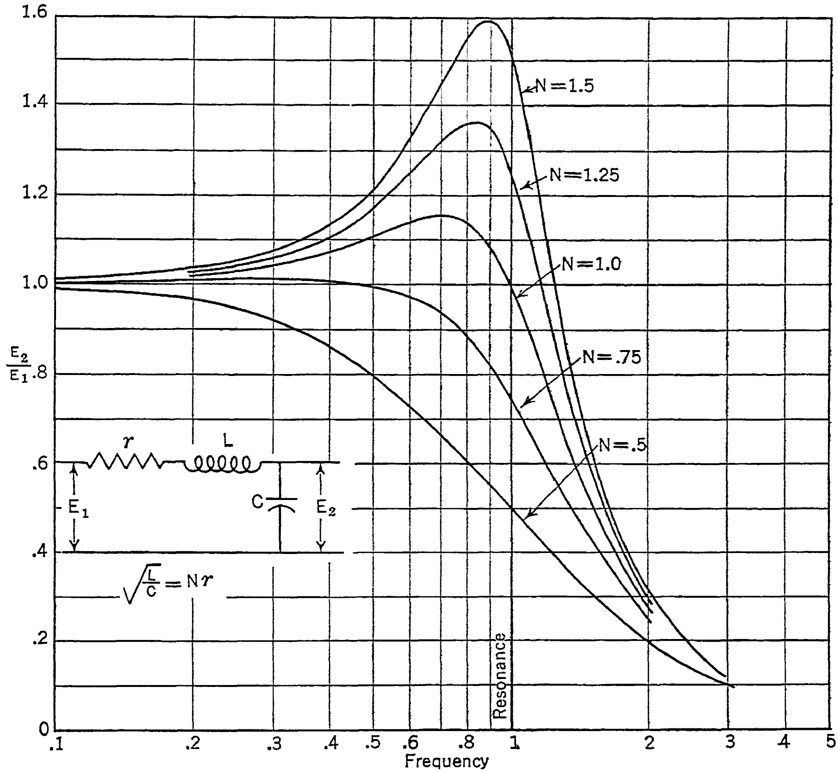


FIG. 16. Leakage Resonance

As an example of what can be expected in the way of turns ratio, assume that all secondary capacitances total 100 $\mu\mu f$ and that the high-frequency response of the transformer is as given by $N = 1$ of Fig. 16, the resonance frequency being chosen as 10,000 cycles. The following step-up ratios are obtainable: with 10,000-ohm generator (triode), 1 to 4; with 500-ohm generator (line), 1 to 17.8; with 100-ohm generator (carbon mike), 1 to 40; with 0.2-ohm generator (ribbon mike), 1 to 890.

PICK-UP AND SHIELDING. Input and interstage transformers often work at very low voltage levels. As a result, voltage induced in the windings by stray magnetic fields may be as large as or larger than the signal voltage. These stray fields are produced by nearby power transformers, rectifier filter reactors, open loops in wires carrying large a-c currents, motor generators, etc. A hum of the frequency of the stray field is introduced into the amplifier.

Correction of hum pick-up should begin at the source, if possible. Reduction of the flux density in power transformers and reactors, by proper design, placing of air gaps of reactors inside of the coils, use of shielding cans around power transformers and reactors, and tight twisting of heavy-current wiring are all helpful.

Removal of the input, or interstage, transformer further from the source of disturbance, and orienting it so that its coil will be at right angles with the coil of the disturbing transformer, are precautions that should be taken when laying out an amplifier.

Input transformers are sometimes made with a two-legged core (see Section 3), half of the primary and half of the secondary being placed on each leg. The two halves of each winding are connected in series, being additive for flux within the core, but subtractive for external fields. Such "hum-bucking" construction is very effective in reducing hum pick-up, if the external field is uniform, so that it acts upon both parts of the transformer equally. A reduction of 40 db in pick-up may be realized.

Shielding of the input, or interstage, transformer is also very effective in reducing pick-up. A drawn nickel-alloy case with a tight-fitting lid will reduce pick-up about 30 db. Two such nickel-alloy shields, one inside the other, separated by a similar copper shield, will reduce pick-up by about 60 db. Three nickel-alloy and two copper shields will give about 90-db reduction of pick-up. Such nested shields are available. See E. B. Harrison, Notes on Transformer Design, *Electronics*, February 1944.

11. DRIVER TRANSFORMER

Function. The Class B operated output stage requires an auxiliary stage of audio amplification called the "driver" stage. The driver transformer couples the plate of the driver tube, usually a triode, to the grids of the Class B amplifier. The function of the driver stage is to supply to the grid circuit of the output stage large positive voltage peaks, which means that the driver stage is required to furnish power. In this respect the driver transformer is similar to the output transformer, but it has additional requirements imposed upon it which make its design more exacting.

TURNS RATIO. The secondary load on a driver transformer varies over a wide range during each half-cycle, from a very high resistance when both grids are negative to a low resistance when either grid is positive. The turns ratio of the transformer must be selected so that the change in load resistance has a negligible effect on driver tube distortion, a condition which is satisfied by using a step-down ratio. The value of the turns ratio is a compromise between distortion and driving power.

FREQUENCY RESPONSE is governed by the same factors as for the output transformer, viz., primary inductance at the low frequencies and leakage inductance at the high frequencies. There is a difference, in that the load on the driver transformer is not a constant resistance but varies during the cycle. Primary inductance should be high enough to give the desired low-frequency response with $r_L = \infty$. Leakage inductance should be low enough to give the desired response when r_L = peak grid voltage swing/peak grid current swing.

LEAKAGE AND DISTORTION. Leakage reactance in the driver transformer is a reactance in series with the grids of the Class B stage. It causes distortion of the grid voltage wave at the higher frequencies. The high-frequency range of a driver transformer is limited by distortion rather than by a falling off of the secondary voltage. Leakage between primary and secondary must therefore be kept to the minimum. It is also important that each half of the secondary be coupled equally to the primary; otherwise the secondary voltage, at high frequencies, will not be the same for both half-cycles, which will produce even harmonics in the grid-voltage wave. The requirements of low leakage and equal coupling are met by using the coil arrangement shown in Fig 17. See T. McLean, An Analysis of Distortion in Class B Audio Amplifiers, *Proc. I.R.E.*, March 1936.

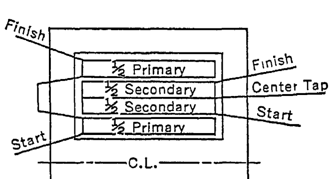


FIG. 17. Winding Arrangement, Class B Driver Transformer

Capacitances of the windings and input capacitance of the tubes have very little effect upon frequency response or distortion. However, they may resonate with the leakage reactance at some superaudible frequency to cause parasitic oscillations of the Class B stage. Such oscillations cannot be suppressed by means of series grid resistors without increasing distortion. It is desirable, therefore, to shunt a small capacitance from each Class B grid to ground.

12. PHYSICAL DESIGN OF AUDIO TRANSFORMERS

Data Required. From the foregoing analyses of various kinds of audio transformers, it is evident that the circuits which the transformer is coupling together must be clearly and completely specified before a design of the transformer is attempted. The first step in designing an audio transformer is to draw a diagram of the circuits, showing the values of generator and load impedance, direct current in any winding, and other pertinent data. Next, the desired constants of the transformer, such as turns ratio, primary inductance, leakage inductance, secondary capacitance, winding resistances, and core-loss resistance are determined from the required performance of the transformer. Methods of finding what constants in the transformer will give desired performance are described above.

DESIGN METHOD. There is no straightforward method of going at the physical design of an audio transformer. The design is carried out by making several trials, each one approaching closer to the desired constants. Procedure is as follows:

1. Assume a core size and a core material. As a starter, a core of EI scrapless laminations (see Section 3, Ferrous-cored Inductors), 1/2 in. to 3/4 in. center leg, stack equal to width of center leg, and silicon-steel material, might be chosen.

2. Calculate the number of primary turns that will give the desired value of primary inductance (Ferrous-cored Inductors).

3. Multiply the number of primary turns by the turns ratio to give the number of secondary turns.

4. Determine the primary and secondary wire sizes, using 500 circular mils per ampere as a first trial, but not using wire smaller than No. 41 AWG. Extremely small and light transformers may employ wire as small as No. 44, but these very small wire sizes should be avoided, if possible, because of breakage when winding. If direct as well as alternating current is present in a winding, the total current rms value will be

$$I_{\text{total}} = \sqrt{I_{\text{dc}}^2 + I_{\text{ac}}^2} \tag{18}$$

5. Lay out the windings. As a rule it is best to calculate the number of turns per layer and the number of layers of each winding rather than to rely on some winding space factor. The arrangement of windings may be fixed by some special requirements of the transformer, as shown in Figs. 13, 14, and 17. Unless there is some reason for doing otherwise, the primary is customarily wound first, with the secondary over it. (See Inductors for details of construction.)

6. If the total calculated build of the coil, including spool, layer insulation, and wrapper, exceeds 90 per cent of the window height, the core is too small, and steps 1, 2, 3, and 5 must be repeated, using either a larger core, a core with a larger window, such as the EE scrapless style, or a core of better magnetic material. If the coil build is far below 90 per cent of the window height, a smaller core should be tried.

7. After working out a core and coil that will fit and will have the desired primary inductance and turns ratio, the other constants should be calculated from the design. These include primary and secondary resistances, core-loss resistance, and leakage inductance. In the case of input and interstage transformers, the distributed capacitance and capacitance to ground of the secondary should also be calculated. In the case of high-level output and modulation transformers, maximum flux density at the lowest frequency (see Inductors), middle-range efficiency, and heating, should also be calculated.

8. Modify the design as required. An inspection of the first trial design will suggest the changes needed. Leakage inductance and secondary capacitance can be varied considerably by changing the arrangement and shape of the windings. Resistance of the windings can be changed by using larger or smaller wire.

LEAKAGE INDUCTANCE depends upon coil geometry. For a two-winding transformer, as shown in Fig. 18, the leakage inductance, referred to the primary, is

$$L_s = \frac{32cN_p^2}{l}\left(a + \frac{d_1 + d_2}{3}\right) \times 10^{-9} \quad \text{henry} \tag{19}$$

in which c = length of a mean turn, a turn halfway between the innermost and outermost layers; l = length of winding, or wire traverse; a = distance between windings, copper to copper; d_1 and d_2 are the build-ups of the two windings, all dimensions being expressed in inches; and N_p is the number of turns of the primary winding. The method of deriving this formula is given by R. R. Lawrence, *Principles of A-C Machinery*. Leakage inductance, referred to the secondary, is given by the same formula, except that N_s^2 is used in place of N_p^2, N_s being the number of turns of the secondary winding.

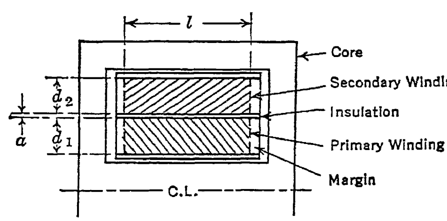


Fig. 18. Cross-section of Transformer Windings

Leakage inductance can be reduced by dividing either the primary or the secondary winding into two sections, placing the other winding between the two sections. One arrangement is shown in Fig. 19*a*, in which the windings are concentrically wound. A second arrangement is shown in Fig. 19*b*, in which the windings are coaxially wound; this construction is termed "pancake" winding. For the two cases:

$$L_e = \frac{8cN_p^2}{l}\left(a_1 + a_2 + \frac{d_1 + d_2 + d_3}{3}\right) \times 10^{-9} \quad \text{henry} \tag{20a}$$

$$L_e = \frac{8cN_p^2}{h}\left(a_1 + a_2 + \frac{d_1 + d_2 + d_3}{3}\right) \times 10^{-9} \quad \text{henry} \tag{20b}$$

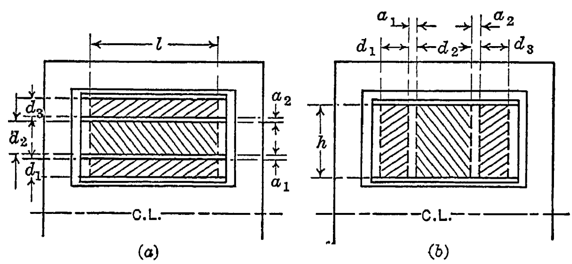


FIG. 19. Interleaved Windings. (*a*) Concentrically Wound. (*b*) Coaxially Wound.

DISTRIBUTED CAPACITANCE of an audio-frequency transformer winding is made up of the layer-to-layer capacitances. The turn-to-turn capacitances are negligible. The equivalent capacitance across a winding is the resultant of the layer-to-layer capacitances, in series. A winding of many layers, therefore, has less distributed capacitance than a winding of few layers. The distributed capacitance, C_d, is

$$C_d = \frac{0.3clk(T-1)}{dT^2} \quad \text{micro-microfarads} \tag{21}$$

in which c = mean length of turn of the winding; l = length of winding, or wire traverse; d = distance between layers, copper to copper, all dimensions expressed in inches; T = number of layers of wire in the winding; and k = average dielectric constant of layer insulation, enamel, and impregnating compounds. For paper-insulated layers, $k = 3$, approximately. See J. H. Morecroft, *Principles of Radio Communication*, 2nd Ed., pp. 233–235.

CAPACITANCE TO GROUND or between windings consists of the capacitances of the inner layer and of the outer layer of a winding to surfaces which are adjacent to them. The capacitances at the ends of a winding are usually negligible. Usually, one winding is wound over another, concentrically, with 10 to 40 mils of insulation between them. Also, the wire traverse is usually the same for both windings. The capacitance between the two windings is the capacitance between two parallel surfaces, of the same area, having a very small separation between them.

It is true that the a-c voltage between the two surfaces is not usually the same at all points, because the a-c voltage across the outer layer of the one winding is usually not the same as that across the inner layer of the other winding. Also, the capacitance calculated between the outer *layer* of the one winding and the inner *layer* of the other is assumed to be from the *finish* of the one winding to the *start* of the other. These are minor errors if the number of layers on each winding is greater than 10.

If the mean circumference of the space between the two windings = c, length of winding = l, and separation of windings = d, all dimensions in inches, and k is the dielectric constant of the insulation between them, the winding-to-winding capacitance, C_w, is

$$C_w = \frac{0.225clk}{d} \quad \text{micro-microfarads} \tag{22}$$

The same formula may be used to compute other capacitances, such as that to core or to shield.

13. AUDIO TRANSFORMER MEASUREMENTS

RESISTANCE. The d-c resistance is usually accurate enough at audio frequencies.

INDUCTANCE AND CAPACITANCE. Most iron-cored transformers for audio and power frequencies resonate at some frequency in the neighborhood of 1000 cycles. This is a parallel resonance of the mutual inductance and the winding capacitances. Measurement of primary or secondary inductance or winding capacitance at 1000 cycles is meaningless. Primary or secondary inductance must be measured at some low frequency, 60 cycles being a convenient one, care being taken to apply an appropriate value of a-c voltage and direct current. Bridges for inductance measurement are described in Section 11.

Capacitance must be measured at some high frequency, such as 4000 cycles or above. Any terminals which are normally at a-c ground potential should be grounded during such measurements. A measurement of capacitance across any winding will include reflected capacitances from other windings. If it is desired to measure the secondary capacitance of an input or interstage transformer, this may be done indirectly by measuring the leakage inductance and the leakage resonance frequency.

LEAKAGE INDUCTANCE is most conveniently measured on a 1000-cycle bridge, short-circuiting one winding and measuring the inductance of the other. This frequency is satisfactory since the mutual inductance and winding capacitances are shorted out, leaving only the leakage inductance and winding resistances as factors in the measurement. If the leakage inductance referred to the primary is to be measured, the secondary is short-circuited and the inductance of the primary is measured. See Fig. 3, in which E_2' would be short-circuited.

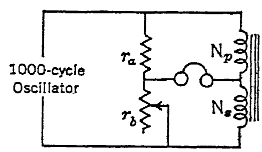


FIG. 20. Turns-ratio Bridge

TURNS RATIO is most accurately and conveniently measured with a bridge as illustrated in Fig. 20. Such a bridge is as accurate as the resistance arms, except when there is a large amount of leakage inductance in the transformer. Polarity is determined at the same time as turns ratio. The windings must be additive in order to obtain a null.

$$\frac{N_s}{N_p} = \frac{r_b}{r_a}$$

CORE-LOSS RESISTANCE, in the middle-frequency range, referred to the primary, is found by measuring the primary impedance, at the self-resonant frequency of the transformer, with no load on the secondary. Figure 21 shows how this impedance may be measured. This impedance is practically equal to the core-loss resistance. Referring to Fig. 3, if kL_1, the primary inductance, is resonant with the total capacitance $C_1 + C_2'$, their combined impedance is infinite, and they have no shunting effect on the circuit. If r_L' and C_L', the load resistance and capacitance, are removed from the transformer, the only shunt left in the circuit is the core-loss resistance r_c. The series elements r_1, the primary winding resistance, and $(1 - k)L_1$, which is half of the leakage inductance, are usually very small as compared with r_c and can be neglected.

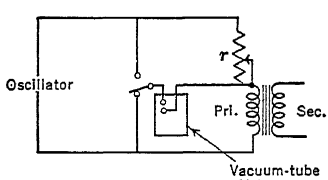


FIG. 21. Measurement of Impedance

FREQUENCY CHARACTERISTIC. An audio oscillator and a suitable voltmeter are required. For measuring input or interstage transformers, the voltmeter must be of the tube type, but for output transformers thermocouple or rectifier-type voltmeters may be used. In any event, the voltmeter should not appreciably load the circuit being measured.

It is essential, when measuring frequency response or phase shift of an audio transformer, that the circuits between which the transformer works, on the primary and secondary sides, be either included in the measurement or that equivalent resistors, capacitors, etc., be used. A measurement of the transformer alone is meaningless. If the actual tube or line on the primary side is not used, an equivalent resistor should be placed in *series* with the primary winding.

Measurements should be made at an a-c voltage level corresponding to actual operating conditions, and with direct current in the windings equal to any unbalanced direct current under actual operating conditions.

14. POWER TRANSFORMER

The function of the power transformer in radio and communication equipment is threefold: to insulate the equipment from the power line, to reduce the line voltage to the voltages required by the tube heater circuits, and to step up the line voltage to energize the anodes of the rectifier that supplies the d-c plate and bias voltages. The secondary windings that supply voltage to heater circuits are usually called "filament" windings, and the secondary winding that supplies voltage to rectifier anodes is termed the "plate" winding. A typical power transformer might have a 115-volt 60-cycle primary, a 600-volt center-tapped plate winding, a 5-volt 2-ampere filament winding for the rectifier heater, and a 6.3-volt 3-ampere filament winding for the other heaters. Variations of this basic pattern are, of course, very numerous, depending upon the voltages and currents needed. Additional secondaries may be required to supply pilot lamps, relays, control motors, etc. In transmitters, separate rectifiers may be used to supply plate and bias voltages, requiring two plate windings. It may be desirable to turn on the heaters of large tubes for a warm-up period before applying the plate voltage, which means that two separate power transformers are required, one to supply the heaters, termed a **filament transformer,** and one to supply the rectifier anodes, termed a **plate transformer.**

Power-line Frequency is 60 cycles per second throughout most of the United States. In some parts of the United States and in many foreign countries 50 cycles is standard. Most power transformers for radio equipment are designed for 50- and 60-cycle operation. In aircraft, 400 or 800 cycles per second is often used.

VOLT-AMPERE RATING. The volt-ampere rating, of any secondary winding is the product of the rms voltage, under load, by the rms current. The total volt-ampere rating of all the secondaries is the sum of the volt-ampere ratings of the individual secondaries. For a filament winding, the volt-ampere rating is simply a-c voltage times a-c current (in amperes). Both voltage and current are of sinusoidal wave form, and the rms value of each is the ordinary a-c effective value. For a plate winding, the rms voltage is the ordinary a-c voltage, since the voltage wave is sinusoidal. However, the current in a plate winding is not sinusoidal. Its rms value depends upon the amount of direct current supplied by the rectifier, the kind of rectifier circuit used, and whether the first element of the rectifier filter is an inductor (choke input) or a capacitor (condenser input). Table 1 gives the ratio of rms current, in the winding, to direct current, for various kinds of rectifier. The factors given for capacitor input are round-number approximations, which are accurate enough for most power-transformer designs. Actually these factors for capacitor input vary widely, depending upon the resistance of the rectifier and the resistance of the d-c load. For analysis of rectifiers see O. H. Schade, Analysis of Rectifier Operation, *Proc. I.R.E.*, July 1943; F. E. Terman, *Radio Engineering*, 2nd Ed., pp. 479–500; R. W. Armstrong, Polyphase Rectification Special Connections, *Proc. I.R.E.*, January 1931.

Table 1. Form Factors for Plate-winding Currents

Type of Rectifier	Heating Current	Volt-drop Current
	I_{rms}/I_{dc}	n
Bridge, inductor input	1.0	1.0
Full-wave, pushpull, inductor input	0.707	0.500
Full-wave, pushpull, capacitor input	1.0	0.707
Bridge, capacitor input	1.5	1.5
Half-wave, capacitor input	2.0	2.0
Voltage-doubler, capacitor input	3.0	3.0

SIZE OF POWER TRANSFORMERS is governed by heating rather than by efficiency in most radio and communications work. Allowable heating limits the flux density in the core and the current density in the windings. If the flux density is fixed, the core cross-section is proportional to the volts per turn. See eq. (26), below. If the current density is fixed, the core window area is proportional to the ampere-turns.

Core cross-section × window area is proportional to volts per turn × ampere-turns.

$$A \times W = p\left(\frac{E}{N} \times NI\right) = p(E \times I) \tag{23}$$

where A is the cross-section of the core, W is the area of the core window, p is a constant of proportionality, N is the number of turns on the secondary of a transformer having but one secondary, E and I are the rms voltage and current of this secondary, and $E \times I$ is the volt-ampere rating of the transformer. When there are several secondaries, the same formula holds, but $E \times I$ is the sum of the volt-ampere ratings of all the secondaries. Applying this formula to the common case of a 60-cycle power transformer, having 40 to 50 deg cent temperature rise, dimensions being in inches,

$$A \times W = 0.024(E \times I) \quad \text{approximately} \tag{24}$$

Power transformers often employ laminations of the EI-scrapless shape (see Section 3, Ferrous-cored Inductors). The core stack is often equal, or nearly so, to the width of the center leg; that is, the center leg has a square cross-section. This core shape has a

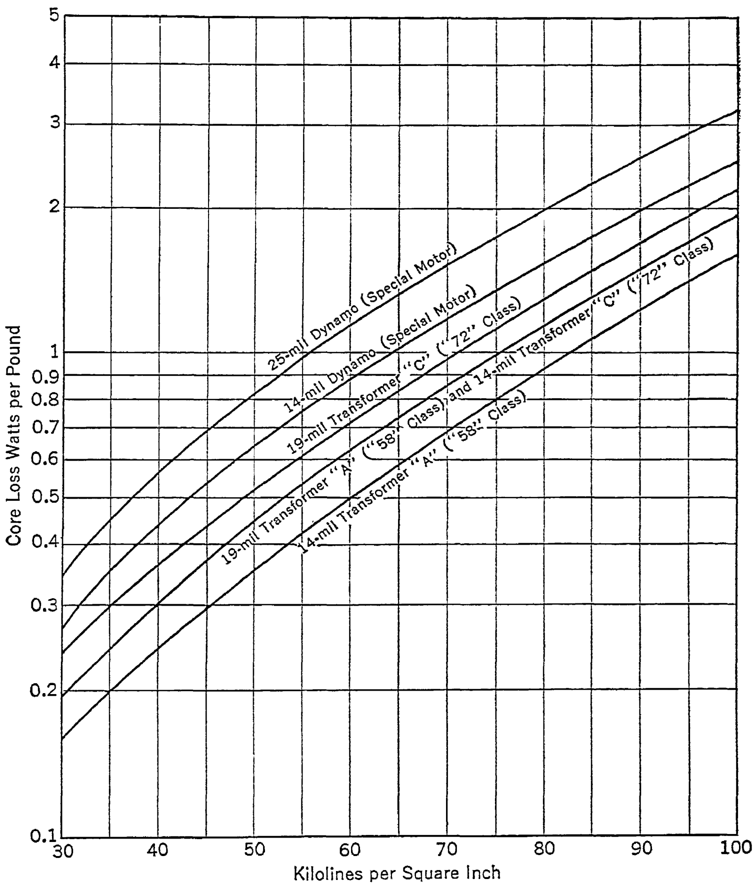


Fig. 22. Core Loss. Silicon Steel at 60 Cycles

definite relationship between window area and core cross-section, $W = 0.75A$. Equation (24), for this case, can be simplified to

$$A = 0.18\sqrt{E \times I} \quad \text{approximately} \tag{25}$$

These formulas are fairly accurate for the average run of power transformers. However, if operating voltages are high, or if the number of secondaries exceeds three, a larger than usual part of the window is taken up by insulation, leaving less space for wire. To allow for this, a somewhat larger core size should be chosen. Often, in such cases, a lamination shape is desirable, having more window area than the EI-scrapless lamination, such as the EE-scrapless style.

CONSTRUCTION OF POWER TRANSFORMERS. Compactness is accomplished by the use of a shell-type core, using laminations of the EI-scrapless or EE-scrapless shapes. Laminations are stacked alternately, or interleaved, one each way or two each way. Core material is almost always silicon steel having 2.5 per cent silicon content, or higher. Lamination thickness may be 14-mil (U. S. gage No. 29), 19-mil (U. S. gage No. 26), or 25-mil (U. S. gage No. 24). The choice of silicon content and lamination thickness is a compromise between cost and core loss, the lower silicon content and greater thickness having the higher core loss. Core loss vs. flux density for several typical grades and thickness of laminations are given in Fig. 22. See Ferrous-cored Inductors, for lamination shapes, kinds of core material, and references.

Coil construction usually follows a conventional pattern, because a particular routine of winding has been found to be most convenient. A single coil, which consists of the various windings, is made, and is placed upon the center leg of the shell-type core. The primary is wound first, over a formed spool of paper or fiber. The winding is layer wound, with paper insulation between layers. Next, an electrostatic shield, consisting of one turn of thin copper, the overlapping ends being insulated from each other, is placed over the primary. The high-voltage or plate winding is wound on next. Up to this point, the winding is done in multiple, ten or more coils being wound at a time. Generally, at this point, the coils are sawed apart. The primary and plate windings ordinarily use wire of sizes too small to be brought out of the windings as leads. So flexible, insulated leads are anchored on top of the plate winding, the wire from the primary and plate windings being brought around the ends of the windings and soldered to these leads. Then the filament windings, which usually consist of a few turns of large wire, are wound on singly, either one over the other or side by side, depending on the number of turns, wire size, and wire traverse required. Generally the wires used for filament windings are sufficiently large and rugged to be extended out of the windings as leads. See Ferrous-cored Inductors. and also H. C. Roters, *Electromagnetic Devices*, Chapter VI.

The coil, or the core and coil together, are baked dry and impregnated with varnish, wax, or asphaltic compound to exclude moisture, strengthen the coil mechanically, and reduce lamination hum.

DESIGN PROCEDURE is relatively straightforward for power transformers, so that design calculations are often made on a standard form sheet, or calculation sheet. Steps are as follows:

1. Determine the rms voltages and currents of the secondary windings and the total volt-ampere rating of all the secondaries.
2. Choose a core size which will satisfy eqs. (24) or (25).
3. Determine, approximately, the primary current. Assume 90 per cent efficiency and 90 per cent power factor as a first approximation. If the plate winding is not center-tapped, the volt-ampere input to the primary will be

$$(\text{Filament volt-amps} + \text{Plate volt-amps}) \div 0.81$$

If the plate winding is center-tapped, the volt-ampere input to the primary will be

$$(\text{Filament volt-amps} + 0.707 \text{ Plate volt-amps}) \div 0.81$$

4. Calculate primary turns, N_p.

$$N_p = \frac{E_p}{4.44BAkf} \times 10^8 \tag{26}$$

in which E_p = primary voltage, f = frequency, B = flux density in lines per square inch, A = cross-sectional area of core in square inches, and k = stacking factor. Flux density is usually around 70 kilolines per square inch, at the nominal primary voltage, for 60-cycle designs.

5. Calculate secondary turns. Assuming a regulation of 10 per cent, as a first trial,

$$N_s = \frac{E_s}{E_p} \times N_p \times 1.10 \tag{27}$$

If the number of turns on the low-voltage filament windings comes out fractional, the turns on all windings should be increased or decreased sufficiently to give each low-voltage winding a whole number of turns.

6. Determine wire sizes of all windings. The rule of 1000 circular mils per ampere is very convenient to use as a first approximation and is usually not far from the correct value arrived at in the final design.
7. Lay out the windings. This is done by calculating the number of turns per layer, the number of layers, and the build of each winding. See above, "Construction of Power Transformers," and Ferrous-cored Inductors. Margins are usually 1/8 in. It is considered good practice to use a double thickness of layer insulation between the two inside layers and between the two outside layers of the high-voltage winding, to provide a factor of safety against transients caused by line surges or switching. The total build of the coil,

including spool, insulation, and outside wrapper, should not exceed 90 per cent of the window height.

Insulation between windings is determined by the operating voltage. The rule of twice normal plus 1000 volts, for test voltage, is commonly used. Insulation must, of course, be able to stand something more than the test voltage. The factor of safety allowed is governed by cost and by the type of service which the radio or communications equipment is required to give. For test voltages of 2500 rms or less, several thicknesses of Kraft paper or two turns of varnished cambric are enough insulation between windings. Margins of $^1/_8$ in. are adequate for creepage. Above 2500 volts rms test, wider margins are necessary as well as better insulation between windings. If, for example, the test voltage is 4000, and a two-to-one safety factor is allowed in the design, the margins must be wide enough to withstand 8000 volts rms creepage. This requires about $^7/_{16}$-in. margins. A point is reached where the margins required for creepage use up too much of the available winding space. Such windings are insulated by wrapping them with half-lapped tape, of varnished cambric, Fiberglas, or other insulating material. A narrow-windowed lamination or side-by-side sections of the winding are employed with very high voltage coils, of 1000 volts or higher, in order to keep the turns per layer and the voltage per layer low.

CALCULATION OF PERFORMANCE. The procedure outlined above will give a design which is usually not very far from a satisfactory final design. It is necessary, however, to calculate the output voltages of the secondaries, the heating, regulation, and efficiency, and then to make minor adjustments in the design as required.

1. Resistance of each winding is calculated from the mean length of turn, in feet, the number of turns, and the resistance of copper wire, of the particular size, in ohms per 1000 ft (see Section 2, for tables). The d-c resistance of the wire is used. An allowance should be made for the fact that ordinary power transformer windings operate at temperatures higher than 20 deg cent, usually in the neighborhood of 80 deg cent. The resistance of copper wire increases about 0.4 per cent per degree centigrade above 20 deg cent. The mean length of turn, m, in feet, is given by

$$m = \frac{2(X + Y) - 8R + \pi(2R + b)}{12} \tag{28}$$

in which X and Y = inside dimensions of the winding, R = inside corner radius, and b = build of the winding, all dimensions being expressed in inches.

2. Copper loss of each winding is the rms current squared $\times$ resistance of that winding. Total copper loss is the sum of the copper losses of the individual windings.

3. Core loss is the product of the watts per pound (Fig. 22) times the weight of the core in pounds. Flux density, for determining watts per pound, is calculated from eq. (26). Actually, the flux density, under load, is somewhat less than given by this equation, owing to voltage drop in the primary winding, but the error is very small. See *Magnetic Circuits and Transformers*, Chapter V.

4. Heating may be calculated from the losses. There is no accurate formula for heating of a power transformer in radio or communications equipment, because so many unpredictable factors are involved. The proximity of other hot objects, such as rectifier tubes or bleeder resistors, the amount of ventilation, the nature of the surfaces toward which the transformer is radiating heat, whether dull or shiny, the kind of finish on the outside of the transformer, and the characteristics of the potting compound, if the transformer is potted, all influence the temperature rise of the transformer. However, a rough rule can be worked out, based upon the Stefan-Boltzmann law for radiation from a black body, which gives an approximate figure for temperature rise and is valuable as a basis for comparing one transformer design against another. The radiation surface of a core and coil is only slightly more than the total surface of the core, including window area. That is, the coil surface adds very little to the radiation surface of the core and coil. E.g., if the outer dimensions of a laminated core are 2 $^1/_2$ by 3 in. and the stack is 1 in., the total core surface is 26 sq. in. This surface is very easy to compute from the core dimensions and may be taken as the radiating surface of the core and coil. Calling this surface A, in square inches, and the total watts loss W, the temperature rise in degrees centigrade, ΔT, is roughly given by

$$\Delta T = \frac{200W}{A} \tag{29}$$

See *Magnetic Circuits and Transformers*, Chapter VIII; Thermal Characteristics of Transformers, V. M. Montsinger, *Gen. Elec. Rev.*, April 1946; and R. Lee, *Electronic Transformers and Circuits*, Wiley, pp. 37–44.

5. Output voltages of the secondaries under load, and the regulation of a transformer, are calculated from the resistances of the various windings. Voltage drop in a transformer,

due to leakage reactance, is very small at 60 cycles and is usually not worth the trouble of computing. An allowance of 1 per cent leakage-reactance drop is usually accurate enough. The full-load voltage of the various secondary windings is obtained as follows:

A. Primary per cent IR drop is

$$\text{Pri. }\% = \frac{I_p \times r_1 \times 100}{E_p} \tag{30}$$

in which I_p = in-phase component of primary current = rms primary current ÷ power factor, r_1 = resistance of primary winding, E_p = impressed primary voltage.

B. No-load secondary voltages = impressed primary voltage × turns ratio, for the various secondary windings.

C. High-voltage winding, full-load voltage is

$$E_2 = \left[E_0 - \left(\frac{\text{Pri. }\% \times E_0}{100} + nI_{\text{dc}}r_2\right)\right] \times 0.99 \tag{31}$$

in which E_0 = no-load voltage of winding, I_{dc} = direct-current rectifier load, n = factor depending on type of rectifier, and r_2 = resistance of high-voltage winding. The 0.99 multiplier is an allowance for leakage-reactance drop. Values of the n factor are given in Table 1. To illustrate where the n factor comes from, consider a full-wave, inductor-input rectifier with winding center-tapped. The direct current flows in one half of the plate winding at a time, so the voltage drop in the plate winding is $0.5 \times r_2 \times I_{\text{dc}}$. The value of n is 0.5.

D. Filament winding, full-load voltage is

$$E_3 = \left[E_0 - \left(\frac{\text{Pri. }\% \times E_0}{100} + I_3r_3\right)\right] \times 0.99 \tag{32}$$

in which E_0 = no-load voltage of the filament winding, I_3 = rms load current, r_3 = resistance of filament winding.

The regulation of each secondary winding, expressed in per cent, is

$$\frac{\text{No-load volts} - \text{Full-load volts}}{\text{Full-load volts}} \times 100$$

Ordinarily, regulation is between 5 and 10 per cent.

6. Efficiency is the ratio, output watts ÷ input watts. The output wattage of any filament winding is simply the product of a-c voltage, under load, by a-c load current. The output wattage of a plate winding is a complicated product of a sinusoidal voltage and a non-sinusoidal current. As an approximation, multiply the rms voltage of the total plate winding by the volt-drop current as given in Table 1. Then the output wattage of the transformer is the sum of the wattages of the various secondaries.

Input wattage is computed by multiplying the *no-load* voltage of each secondary by its rms load current if a filament winding, or by the volt-drop current if a plate winding, and adding the core-loss watts. The efficiency of a typical power transformer is about 90 per cent.

15. VIBRATOR TRANSFORMER

Function. The vibrator transformer is used in radio and communications equipment when the source of power is a battery instead of an a-c line. This occurs in mobile applications, such as automobile, aircraft, and railroad. The vibrator transformer takes the place of the usual power transformer, supplying the proper voltages to rectifier anodes and to heaters.

The center tap of the primary winding is connected permanently to one end of the battery. The other end of the battery is connected alternately to one end or the other of the primary winding by means of contacts on a vibrating reed. The frequency of the reed is usually between 100 and 200 cycles per second, although frequencies as high as 400 cycles are used. This applies an alternating voltage, of square wave form, to the primary of the transformer. The frequency of this alternating voltage is, of course, the same as the vibration frequency of the reed. The transformer is able to step this primary voltage up or down, as required, in much the same manner as an ordinary a-c voltage.

An appreciable time is required for the reed to move from one side to the other, so that the contacts are closed, one way or the other, only about 80 per cent of the time. The transformer is connected across the battery only that percentage of the time. The portion of the time that the vibrator contacts are closed is called the **time efficiency** of the vibrator.

A capacitor, called a "buffer" capacitor, must be connected across either the primary or the secondary winding. Usually it is placed across the highest voltage winding because a smaller value of capacitance will suffice there. The buffer capacitor is an essential part of the transformer. When the correct value of buffer is used, it gives a smooth cross-over of the transformer voltage during the time interval when the vibrator contacts are open. In so doing, it prevents sparking and high-frequency transients at the break and at the make.

Often, heater circuits are connected directly to the battery, leaving only rectifier anode voltage to be supplied by the vibrator transformer. The rectifier winding may be connected to vacuum-tube rectifiers in the same manner as the plate winding of an ordinary power transformer. Or rectification may be accomplished by using a second pair of contacts on the vibrator which switch the load back and forth across the secondary in synchronism with the switching of the battery across the primary. Figure 23 shows the circuit of a synchronous vibrator transformer.

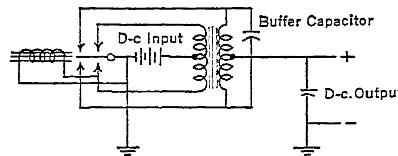


FIG. 23. Circuit of Synchronous Vibrator

DESIGN. (Rectifier secondary only.) A. Core size is much larger than for an ordinary power transformer of the same frequency and giving the same d-c voltage and current. As battery voltage varies widely, flux density must be kept low. The primary winding is almost twice as bulky as that of an ordinary transformer. The current is flowing only part of the time, and so larger wire is required for the same average current. On the other hand, if the frequency of the vibrator is much higher than 60 cycles, e.g., 150 cycles, the higher frequency will reduce the size of the vibrator transformer. It will still be somewhat larger than an equivalent 60-cycle power transformer.

B. Flux density and primary turns. The flux density in a vibrator-transformer core is given by

$$B_{\text{max}} = \frac{(1+p)E_1}{4N_pAkf} \times 10^8 \quad \text{lines per square inch} \tag{33}$$

in which E_1 = normal voltage of the battery, p = time efficiency, N_p = total primary turns, A = cross-section of core in square inches, k = stacking factor of core, and f = vibrator frequency. Flux density should be kept below 40,000 lines per square inch at normal battery voltage. A still lower figure may be desirable at higher frequencies because of core loss.

C. Secondary turns. As a first approximation, use

$$N_s = 1.33 \times N_p \times \frac{\text{Desired d-c output volts}}{\text{Battery volts}} \tag{34}$$

D. Current and wire size. The rms current in the secondary winding is

$$I_s = \frac{0.707 \times I_{\text{dc}}}{\sqrt{p}} \tag{35a}$$

The rms current in the primary winding is

$$I_p = \frac{0.707 \times I_{\text{dc}}}{\sqrt{p}} \times \frac{N_s}{N_p} \tag{35b}$$

in which I_{dc} = d-c load current of the secondary. Wire sizes should be chosen to have about 800 circular mils per ampere. Usually the primary has relatively few turns because the battery voltage is a low voltage. The primary wire size should be chosen so as to give 2, 4, 6, or 8 even layers. This will place the center tap at the end of a layer, which is convenient for multiple winding.

E. Output voltage. The total resistance of the transformer and associated circuits, referred to the working half of the secondary, is

$$r_T = \left(\frac{r_1}{2} + r_v + r_L\right)\left(\frac{N_s}{N_p}\right)^2 + \frac{r_2}{2} + r_3 \quad \text{ohms} \tag{36}$$

in which the r symbols indicate resistances, as follows: r_1 = total primary, r_v = vibrator contact, r_L = battery leads, r_2 = total secondary, and r_3 = rectifier tube, if any.

During the time that the vibrator contacts are closed, the secondary current = I_{dc}/p. Then, the IR drop expressed in volts on the secondary is $r_T \times I_{\text{dc}}/p$.

The no-load d-c voltage across the first filter capacitor is

$$E_0 = aE_1 \frac{N_s}{N_p} \tag{37}$$

a being a constant which takes care of imperfect contacts, primary and secondary contacts not exactly synchronized, etc., = 0.94 for synchronous rectifier, 0.98 for tube rectifier.

Then the full-load d-c voltage across the first filter capacitor is

$$E_2 = E_0 - \frac{r_T I_{dc}}{p} \tag{38}$$

See T. T. Short and J. P. Coughlin, Try the Inverter Transformer, *Elec. Mfg.*, June 1946; F. E. Terman, *Radio Engineering*, pp. 500–501; Mallory Vibrator Data Book.

16. PULSE TRANSFORMER

In radar work, iron-cored transformers are used, which have voltage impressed upon them for very short periods. These pulses are repeated at regular intervals, the time interval between pulses being perhaps 1000 times the pulse duration. Analysis of such pulse transformers cannot be made on a basis of Fourier analysis of the wave form because of the relatively great time interval between pulses. Each pulse is a separate transient, the effect of which dies out before the next pulse, except for core magnetization.

If a voltage is applied suddenly to the primary, through a generator resistance such as the plate resistance of a modulator tube, then is held constant for a short time interval, and then is suddenly removed, the input voltage will be of square pulse shape. The transformer will step up, or step down, this voltage, in accordance with its turns ratio, but the output voltage will not faithfully follow the square pulse shape. An appreciable time is required for the secondary voltage to build up from zero to its maximum value. This "rise time" is caused by the leakage inductance and capacitance of the transformer. During the time that the input voltage is being held constant, the output voltage will be falling off, the amount of drop from a constant voltage value being inversely proportional to the primary inductance. When the input voltage is removed, the secondary voltage does not drop instantly to zero but drags out through several damped oscillations caused by the discharge of magnetic energy stored in the core through the winding capacitances and load resistance. Figure 24 illustrates the kind of deformation of a pulse that occurs when it is passed through a transformer. See R. Lee, Iron-core Components in Pulse Amplifiers, *Electronics*, August 1943, and R. Lee, *Electronic Transformers and Circuits*, Wiley, Chapter IX.

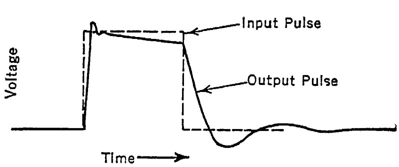


FIG. 24. Deforming of Pulse Shape by Transformer

The pulse transformer has a number of unique features. Because of the very low-duty cycle, tremendous pulse power can be handled by a very small transformer. The requirement of low leakage inductance is met by close spacing between the primary and secondary windings and by making the primary layer or layers exactly the same length as those of the secondary, even to the extent of winding several wires in parallel on the low-voltage winding. If a winding requires more than one layer, the conventional forward-and-back method of winding is not used; layers are all started from the same end to minimize capacitance between layers.

The small physical size and the relatively few turns which are essential to obtain low leakage inductance together with very high pulse voltages make voltage gradients necessary that are unheard of in ordinary transformer design. For example, 200 or more volts per turn is not uncommon. This is accomplished by using Formvar-coated wire and impregnating the transformer with transformer oil under a very high vacuum to remove every trace of air.

High primary inductance is required, to keep the drop of voltage during the pulse to the minimum. A new conception of incremental permeability is necessary in calculating the inductance. All the pulses are of the same polarity; consequently the core is left in a partially magnetized condition, owing to remanence. The hysteresis loop described by the core has this remanent point as one of its ends. The shape of the loop is largely controlled by eddy-current loss when the pulses are very short, such as of 1 microsecond duration. Under these conditions incremental permeability is much lower than when a core is operated with alternating current, and core material having low eddy-current loss is desirable. Ribbon cores of very thin silicon-steel ribbon, 1 to 3 mils thick, are very satisfactory for pulse transformer cores. Their incremental permeability under microsecond pulse conditions is about 300.

ELECTRIC WAVE FILTERS

By A. J. Grossman

A wave filter is a device for separating waves characterized by a difference in frequency. The general purpose of an electric wave filter is to separate sinusoidal electrical currents of different frequencies. Ideally, a filter transmits freely the currents of all frequencies lying within a specified range and excludes currents of all other frequencies. It may be used to transmit the intelligence contained in a certain band and exclude adjacent steady-state interference; combinations of filters may divide a wide frequency band into a number of relatively narrow channels or may direct selected bands from one transmission path into two or more different paths. Except in the simplest forms, a filter is a composite network made up of several sections connected in tandem. Each section consists of simple arrangements of two-terminal reactance networks. These reactances are provided by combinations of ordinary coils and condensers, crystals, coaxial lines, and/or wave guides.

The point of view developed by Bode, Campbell, Cauer, Foster, and others for the analysis of a network is to regard the combination of inductances and capacitances as a system excited by a vibratory disturbance to which the methods of particle dynamics can be applied. It is convenient to express the analysis in terms derived from the classical theory of wave propagation in continuous media. These terms are the image impedance and the image transfer constant. In this terminology a filter is described as a system having the following idealized properties. Signals lying within a preassigned frequency band are transmitted without reduction in amplitude. This band is bounded by cutoff frequencies at which there is an abrupt transition from free transmission to attenuation. The attenuation increases more or less rapidly with frequency as the departure from a cutoff increases. Concomitantly, the impedance is a pure resistance in the pass band and changes abruptly at a cutoff to a pure reactance.

These idealized characteristics are approached in an actual filter inserted between resistance terminations. The insertion loss for frequencies in the pass band remote from the cutoff is essentially nil (apart from the effect due to dissipation in the components). As the cutoff is approached the loss increases. The transition from the theoretical pass band to the attenuating band is smooth. This transition interval can be made extremely narrow in an elaborate design. The cutoff is not a frequency at which there is an abrupt change from zero attenuation to a large value; it may mark the point at which there is a rapid change from a small to a large value of attenuation, but at a finite rate. This departure from the idealized characteristics arises from the fact that the filter is not terminated in its image impedances. In simple cases, the image impedance varies considerably with frequency, and, consequently, the input impedance of the filter has a non-constant resistance component and an associated reactance component. By careful design the image impedance may be maintained nearly constant over almost all of the pass band. Nevertheless, there is still the smooth, but rapid, transition of the input impedance from a predominantly resistive characteristic in the pass band to a predominantly reactive characteristic in the attenuating band. The usual types of filters are low-pass, high-pass, and band-pass.

17. INTRODUCTION

A general filter network may be represented by a box, as in Fig. 1, having two pairs of accessible terminals. The performance of this network is described in terms of its image impedances and image transfer constant. The image impedances are defined as those impedances with which the network must be terminated so that there will not be reflections at the junctions 1–1′ and 2–2′. That is, when the terminating impedance, Z_{I_1}, is connected across 1–1′, then the impedance measured at the 2–2′ terminals is Z_{I_2}; and, similarly, when Z_{I_2} is connected across 2–2′, the impedance measured at 1–1′ is Z_{I_1}.

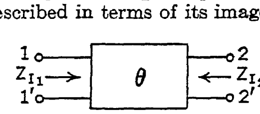
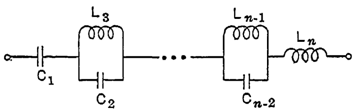
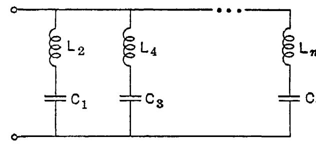


FIG. 1. A General Filter Network

The image transfer constant, θ, is defined to be equal to one-half the natural logarithm of the ratio of the volt-amperes flowing into the network to the volt-amperes flowing

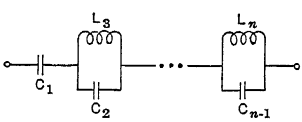
(1) $$Z = \frac{H(\omega_1^2 - \omega^2)(\omega_3^2 - \omega^2)\cdots(\omega_{n-1}^2 - \omega^2)}{j\omega(\omega_2^2 - \omega^2)(\omega_4^2 - \omega^2)\cdots(\omega_{n-2}^2 - \omega^2)}; \quad n = \text{even integer}$$

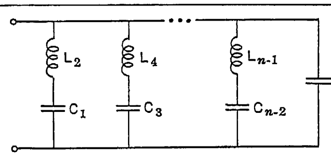




$C_{2k} = \left[\frac{j\omega}{(\omega_{2k}^2 - \omega^2)Z}\right]_{\omega=\omega_{2k}}; \ k = 1, 2, \cdots, \frac{n-2}{2}$	$L_{2k} = \left[\frac{j\omega Z}{(\omega_{2k-1}^2 - \omega^2)}\right]_{\omega=\omega_{2k-1}}; \ k = 1, 2, \cdots, \frac{n}{2}$
$L_{2k+1} = \frac{1}{\omega_{2k}^2 C_{2k}}$	$C_{2k-1} = \frac{1}{\omega_{2k-1}^2 L_{2k}}$
$C_1 = \frac{1}{H} \cdot \left[\frac{\omega_2\omega_4 \cdots \omega_{n-2}}{\omega_1\omega_3 \cdots \omega_{n-1}}\right]^2$	
$L_n = H$	

(2) $$Z = \frac{H(\omega_1^2 - \omega^2)(\omega_3^2 - \omega^2)\cdots(\omega_{n-2}^2 - \omega^2)}{j\omega(\omega_2^2 - \omega^2)(\omega_4^2 - \omega^2)\cdots(\omega_{n-1}^2 - \omega^2)}; \quad n = \text{odd integer}$$





$C_{2k} = \left[\frac{j\omega}{(\omega_{2k}^2 - \omega^2)Z}\right]_{\omega=\omega_{2k}}; \ k = 1, 2, \cdots, \frac{n-1}{2}$	$L_{2k} = \left[\frac{j\omega Z}{(\omega_{2k-1}^2 - \omega^2)}\right]_{\omega=\omega_{2k-1}}; \ k = 1, 2, \cdots, \frac{n-1}{2}$
$L_{2k+1} = \frac{1}{\omega_{2k}^2 C_{2k}}$	$C_{2k-1} = \frac{1}{\omega_{2k-1}^2 L_{2k}}$
$C_1 = \frac{1}{H} \cdot \left[\frac{\omega_2\omega_4 \cdots \omega_{n-1}}{\omega_1\omega_3 \cdots \omega_{n-2}}\right]^2$	$C_n = \frac{1}{H}$

FIG. 2. Design Information for

out of the network when the network is terminated in its image impedances. These quantities are an exact measure of the performance of the filter only if the actual terminations are equal to the image impedances. In a practical design, account must be taken of reflection and interaction effects arising from the mismatch between the terminating impedances of the filter and its image impedances. These effects may be evaluated by the method described in Section 5. They will not be considered in detail here. It will be assumed that the performance of the filter is described by its image transfer constant.

By writing the mesh or nodal equations for the network, it may be shown that:

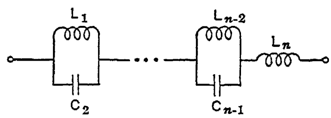
$$Z_{I_1} = \sqrt{Z_{s1}Z_{01}} \tag{1}$$

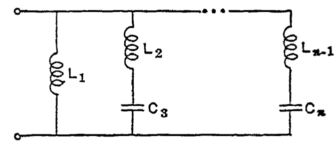
$$Z_{I_2} = \sqrt{Z_{s2}Z_{02}} \tag{2}$$

$$\tanh\theta = \sqrt{\frac{Z_{s1}}{Z_{01}}} = \sqrt{\frac{Z_{s2}}{Z_{02}}} \tag{3}$$

where Z_{s1} is the impedance measured at the 1–1′ terminals when a short circuit is placed across the 2–2′ terminals, and Z_{01} is the impedance measured at the 1–1′ terminals when the 2–2′ terminals are open. The short- and open-circuit impedances Z_{s2} and Z_{02} at the 2–2′ terminals are measured similarly.

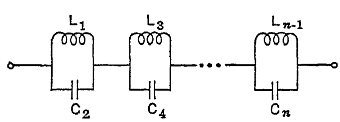
(3) $$Z = j\frac{H\omega(\omega_2^2-\omega^2)(\omega_4^2-\omega^2)\cdots(\omega_{n-1}^2-\omega^2)}{(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)\cdots(\omega_{n-2}^2-\omega^2)};\quad n = \text{odd integer}$$

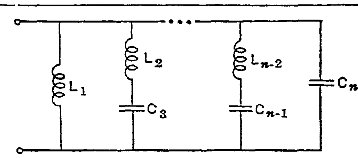




Series type	Parallel type
$C_{2k} = \left[\frac{j\omega}{(\omega_{2k-1}^2-\omega^2)Z}\right]_{\omega=\omega_{2k-1}}$; $k = 1, 2, \cdots, \frac{n-1}{2}$	$L_{2k} = \left[\frac{j\omega Z}{(\omega_{2k}^2-\omega^2)}\right]_{\omega=\omega_{2k}}$; $k = 1, 2, \cdots, \frac{n-1}{2}$
$L_{2k-1} = \frac{1}{\omega_{2k-1}^2 C_{2k}}$	$C_{2k+1} = \frac{1}{\omega_{2k}^2 L_{2k}}$
$L_n = H$	$L_1 = H\cdot\left[\frac{\omega_2\omega_4\cdots\omega_{n-1}}{\omega_1\omega_3\cdots\omega_{n-2}}\right]^2$

(4) $$Z = j\frac{H\omega(\omega_2^2-\omega^2)(\omega_4^2-\omega^2)\cdots(\omega_{n-2}^2-\omega^2)}{(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)\cdots(\omega_{n-1}^2-\omega^2)};\quad n = \text{even integer}$$





Series type	Parallel type
$C_{2k} = \left[\frac{j\omega}{(\omega_{2k-1}^2-\omega^2)Z}\right]_{\omega=\omega_{2k-1}}$; $k = 1, 2, \cdots, \frac{n}{2}$	$L_{2k} = \left[\frac{j\omega Z}{(\omega_{2k}^2-\omega^2)}\right]_{\omega=\omega_{2k}}$; $k = 1, 2, \cdots, \frac{n-2}{2}$
$L_{2k-1} = \frac{1}{\omega_{2k-1}^2 C_{2k}}$	$C_{2k+1} = \frac{1}{\omega_{2k}^2 L_{2k}}$
	$L_1 = H\cdot\left[\frac{\omega_2\omega_4\cdots\omega_{n-2}}{\omega_1\omega_3\cdots\omega_{n-1}}\right]^2$
	$C_n = \frac{1}{H}$

Two-terminal Reactive Networks

The short- and open-circuit impedances are driving point impedances of a purely reactive network. The requirements on such an impedance are (by Foster's theorem):

It is an odd rational function of the frequency, $\omega/2\pi$, which is completely determined, except for a constant factor, H, by assigning the resonant and antiresonant frequencies, subject to the condition that they alternate and include both zero and infinity. Such an impedance function may be physically realized by several canonical structures, among them a combination of antiresonant circuits connected in series and a combination of resonant circuits connected in parallel.

Figure 2 illustrates the four possible reactance functions which are distinguished by their behavior at zero and infinite frequency. The series-type networks are specified in the left-hand column, and the parallel type in the right-hand. The number of elements in each configuration is the minimum, and equal to the number of critical frequencies plus 1.

The operations required to evaluate an expression such as

$$C_{2k} = \left[\frac{j\omega}{(\omega_{2k}^2-\omega^2)Z}\right]_{\omega=\omega_{2k}}$$

are to be performed in the following sequence: (1) multiply $\frac{1}{Z}$ by $\frac{j\omega}{\omega_{2k}^2-\omega^2}$; (2) cancel the common factor $(\omega_{2k}^2-\omega^2)$; (3) replace ω by ω_{2k}.

Illustration.

$$Z = \frac{H(\omega_1^2 - \omega^2)(\omega_3^2 - \omega^2)}{j\omega(\omega_2^2 - \omega^2)(\omega_4^2 - \omega_2)}$$

This is realizable with the second pair of networks. For the series type:

$$C_2 = \frac{-\omega_2^2(\omega_4^2 - \omega_2^2)}{H(\omega_1^2 - \omega_2^2)(\omega_3^2 - \omega_2^2)}$$

$$C_4 = \frac{-\omega_4^2(\omega_2^2 - \omega_4^2)}{H(\omega_1^2 - \omega_4^2)(\omega_3^2 - \omega_4^2)}$$

$$L_3 = \frac{1}{\omega_2^2 C_2}$$

$$L_5 = \frac{1}{\omega_4^2 C_4}$$

$$C_1 = \frac{1}{H}\left(\frac{\omega_2\omega_4}{\omega_1\omega_3}\right)^2$$

For the parallel type:

$$L_2 = \frac{H(\omega_3^2 - \omega_1^2)}{(\omega_2^2 - \omega_1^2)(\omega_4^2 - \omega_1^2)}$$

$$L_4 = \frac{H(\omega_1^2 - \omega_3^2)}{(\omega_2^2 - \omega_3^2)(\omega_4^2 - \omega_3^2)}$$

$$C_1 = \frac{1}{\omega_1^2 L_2}$$

$$C_3 = \frac{1}{\omega_3^2 L_4}$$

$$C_5 = \frac{1}{H}$$

18. PROPERTIES OF THE IMAGE PARAMETERS

GENERAL CONDITIONS. In the pass band of a filter, the image impedances are positive real quantities, or resistances; the image transfer constant is pure imaginary, which signifies phase shift with zero attenuation. In the attenuating region, the image impedances are pure imaginary, or reactances; the transfer constant has a positive real component, which signifies attenuation.

COINCIDENCE CONDITIONS. According to eqs. (1) and (3), the image impedance at one end of the network, and the transfer function, depend only on the short- and open-circuit impedances measured at that end. Typical expressions for these impedances are:

$$Z_{s1} = \frac{H_1(a_1^2 - \omega^2)(a_3^2 - \omega^2)\cdots(a_m^2 - \omega^2)}{j\omega(a_2^2 - \omega^2)(a_4^2 - \omega^2)\cdots(a_{m-1}^2 - \omega^2)} \tag{4}$$

$$Z_{01} = \frac{H_2(b_1^2 - \omega^2)(b_3^2 - \omega^2)\cdots(b_k^2 - \omega^2)}{j\omega(b_2^2 - \omega^2)(b_4^2 - \omega^2)\cdots(b_{k-1}^2 - \omega^2)} \tag{5}$$

where a_1, a_3, a_5, $\cdots$ are values of the angular frequency at which the short-circuited network is resonant, and a_2, a_4, $\cdots$ are values of the angular frequency at which it is antiresonant. Similarly, the b's with odd subscripts are the resonant, and the b's with even subscripts the antiresonant, angular frequencies of the network when the far end is open-circuited.

The expressions for the image impedance and transfer constant obtained with these impedance functions are:

$$Z_{I_1}^2 = \frac{H_1H_2}{-\omega^2}\cdot\frac{(a_1^2 - \omega^2)\cdots(a_m^2 - \omega^2)}{(a_2^2 - \omega^2)\cdots(a_{m-1}^2 - \omega^2)}\cdot\frac{(b_1^2 - \omega^2)\cdots(b_k^2 - \omega^2)}{(b_2^2 - \omega^2)\cdots(b_{k-1}^2 - \omega^2)} \tag{6}$$

$$\tanh^2\theta = \frac{H_1}{H_2}\cdot\frac{(a_1^2 - \omega^2)\cdots(a_m^2 - \omega^2)}{(a_2^2 - \omega^2)\cdots(a_{m-1}^2 - \omega^2)}\cdot\frac{(b_2^2 - \omega^2)\cdots(b_{k-1}^2 - \omega^2)}{(b_1^2 - \omega^2)\cdots(b_k^2 - \omega^2)} \tag{7}$$

The statement of the general conditions on the image parameters indicates that the network transmits freely when $Z_{I_1}^2$ is positive and $\tanh^2\theta$ is negative; it attenuates those frequencies for which $Z_{I_1}^2$ is negative and $\tanh^2\theta$ is positive. In general, the expressions (6) and (7) change sign as the frequency passes through each a and b, and the network has a multitude of pass and attenuating bands. In order that the network be a filter which

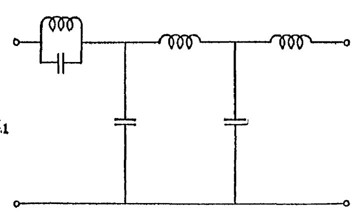
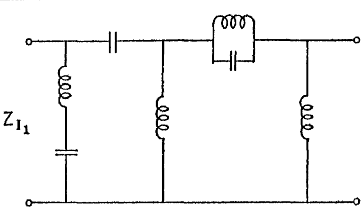
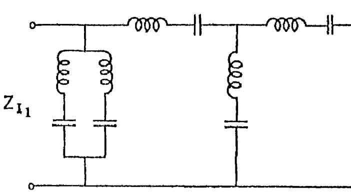
Filter Type	Low-pass	High-pass	Band-pass
Frequency pattern	Z_{s1}: o x o x o x o Z_{01}: x o x o · x o $\tanh\theta$: o x o x $\boxed{\text{o}}$ · · Z_{I_1}: · · · · $\boxed{\text{o}}$ x o $0\quad \omega_1\quad \omega_2\quad \omega_3\quad \omega_4\quad \omega_5\quad \infty$	Z_{s1}: x o · x o x o Z_{01}: x o x o x o x $\tanh\theta$: · · $\boxed{\text{o}}$ x o x o Z_{I_1}: x o $\boxed{\text{x}}$ · · · · $0\quad \omega_1\quad \omega_2\quad \omega_3\quad \omega_4\quad \omega_5\quad \infty$	Z_{s1}: x o x o x o x o x Z_{01}: x o · x o x · o x $\tanh\theta$: · · $\boxed{\text{x}}$ o x o $\boxed{\text{x}}$ · · Z_{I_1}: x o $\boxed{\text{x}}$ · · · $\boxed{\text{x}}$ o x $0\quad \omega_1\quad \omega_2\quad \omega_3\quad \omega_4\quad \omega_5\quad \omega_6\quad \omega_7\quad \infty$
Z_{s1}	$\dfrac{H_1 j\omega(\omega_2^2-\omega^2)(\omega_4^2-\omega^2)}{(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}$	$\dfrac{H_1(\omega_1^2-\omega^2)(\omega_4^2-\omega^2)}{j\omega(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}$	$\dfrac{H_1(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)(\omega_7^2-\omega^2)}{j\omega(\omega_2^2-\omega^2)(\omega_4^2-\omega^2)(\omega_6^2-\omega^2)}$
Z_{01}	$\dfrac{H_2(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)}{j\omega(\omega_2^2-\omega^2)(\omega_5^2-\omega^2)}$	$\dfrac{H_2(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}{j\omega(\omega_2^2-\omega^2)(\omega_4^2-\omega^2)}$	$\dfrac{H_2(\omega_1^2-\omega^2)(\omega_4^2-\omega^2)(\omega_7^2-\omega^2)}{j\omega(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}$
$\tanh\theta$	$\sqrt{\dfrac{H_1}{H_2}}\cdot\dfrac{j\omega(\omega_2^2-\omega^2)\sqrt{\omega_4^2-\omega^2}}{(\omega_1^2-\omega^2)(\omega_3^2-\omega^2)}$	$\sqrt{\dfrac{H_1}{H_2}}\cdot\dfrac{\sqrt{\omega_2^2-\omega^2}(\omega_4^2-\omega^2)}{(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}$	$\sqrt{\dfrac{H_1}{H_2}}\cdot\dfrac{(\omega_3^2-\omega^2)(\omega_5^2-\omega^2)}{\sqrt{\omega_2^2-\omega^2}(\omega_4^2-\omega^2)\sqrt{\omega_6^2-\omega^2}}$
Z_{I_1}	$\sqrt{H_1H_2}\cdot\dfrac{\sqrt{\omega_4^2-\omega^2}}{(\omega_5^2-\omega^2)}$	$\sqrt{H_1H_2}\cdot\dfrac{(\omega_1^2-\omega^2)}{j\omega\sqrt{\omega_2^2-\omega^2}}$	$\sqrt{H_1H_2}\cdot\dfrac{(\omega_1^2-\omega^2)(\omega_7^2-\omega^2)}{j\omega\sqrt{(\omega_2^2-\omega^2)(\omega_6^2-\omega^2)}}$
Possible network	Z_{I_1} 	Z_{I_1} 	Z_{I_1} 

FIG. 3. Illustrative Expressions for the Short- and Open-circuit Impedance, Image Impedance, and Image Transfer Functions

passes one continuous band of frequencies and attenuates all other bands, it is necessary to place certain evident restrictions on the critical frequencies of its short- and open-circuit impedances. If, for example, $a_1 = b_1 = \omega_1$, the factors $(a_1^2 - \omega^2)$ and $(b_1^2 - \omega^2)$ cancel each other in $\tanh^2 \theta$ but form the factor $(\omega_1^2 - \omega^2)^2$ in $Z_{I_1}^2$. In either event, as ω passes through the value ω_1 the sign of the expression does not change. It follows that the elements of the network must be so chosen that all the a's are equal to, or coincide with, b's except possibly in two cases. These exceptional cases correspond to cutoff frequencies at which the network changes from a condition of free transmission to attenuation, or vice versa. These transition points may be two a's or two b's or one a and one b.

For the typical short- and open-circuit impedances under consideration, it is seen from (6) and (7) that $Z_{I_1}^2$ is negative and $\tanh^2 \theta$ is positive at zero frequency. That is, the filter attenuates zero frequency. This condition will continue over a frequency band provided that $a_1 = b_1$, $a_2 = b_2$, or, in general, $a_j = b_j$. The attenuation band is terminated at a cutoff frequency by locating an a at that frequency without a corresponding b, or vice versa. In either case, if a continuous pass band is to extend beyond the cutoff, it is necessary that the subsequent coincidences of critical frequencies be of a type specified by the formula $a_j = b_{j\pm 1}$, since the sequence either of a's or of b's has lost a step at the cutoff. The pass band may continue to infinite frequency or it may be terminated at a second cutoff. This cutoff may be specified by either an a or a b. Then the coincidences in the attenuating band above this cutoff are given by the formula $a_j = b_j$ or $a_j = b_{j\pm 2}$. The first relation holds if one cutoff is an a and the other a b; the second, if both cutoffs are a's or b's.

Inspection of eqs. (6) and (7) leads to the following conclusions: (*a*) coincidences of the type $a_j = b_j$ and $a_j = b_{j\pm 2}$ produce double zeros or poles in $Z_{I_1}^2$ but cancel out in $\tanh^2 \theta$; (*b*) coincidences of the type $a_j = b_{j\pm 1}$ produce double zeros or poles in $\tanh^2 \theta$ but cancel out in $Z_{I_1}^2$; (*c*) critical frequency coincidences of the first type produce image-impedance-controlling factors; those of the second type produce transfer-constant-controlling factors.

SUMMARY OF PROPERTIES OF THE IMAGE PARAMETERS.

1. $\tanh^2 \theta$ is the ratio of two impedance functions of a reactive network, and Z_I^2 is the product of two such functions. $\tanh^2 \theta$ and Z_I^2 contain only double zeros and poles except possibly for two zeros or poles, which are simple.

2. The simple zeros or poles represent cutoff frequencies and occur at positive values of ω^2. They are the same for the two expressions except for the possibility that either or both may be a zero in one expression and a pole in the other.

3. The zeros and poles of $\tanh^2 \theta$ alternate with each other in each continuous pass band; the zeros and poles of Z_I^2 alternate with each other in each continuous attenuating band. The step between bands may interrupt the alternation since the cutoffs may both be zeros or both poles.

Illustration. Examples of expressions for the short- and open-circuit impedances and the image parameters are given in Fig. 3. The frequency pattern is a convenient schematic representation of these functions. It is a plot of the location of the zeros and poles of a function in the frequency scale. The conventions are: circles denote zeros of the function (or resonances of the impedance); crosses denote poles of the function (or antiresonances of the impedance); squares represent cutoffs. These diagrams serve to illustrate the properties of the image parameters summarized above. The network configurations shown in the table are "possible" in the sense that they will be obtained for an appropriate choice of the multipliers H_1 and H_2 and the critical frequencies of the short- and open-circuit impedances.

19. OPEN-CIRCUIT TRANSFER IMPEDANCE

Two important filter theorems, due to H. W. Bode, can be derived from the properties of the open-circuit transfer impedance. This impedance, designated Z_{012}, is the ratio of the voltage appearing across the open output terminals of the network to the current fed into the input terminals. For a purely reactive network, this impedance function has real coefficients and is imaginary at real values of frequency. It is expressed in terms of the short- and open-circuit impedances by:

$$Z_{012}^2 = Z_{01}Z_{02} - Z_{01}Z_{02} = Z_{01}Z_{02} - Z_{01}Z_{02} \tag{8}$$

By use of eqs. (1)–(3), this may be written:

$$Z_{012}^2 = (Z_{I_1}Z_{I_2})\left(\frac{1}{\tanh^2 \theta} - 1\right) \tag{9}$$

Since Z_{012} is imaginary along the real frequency axis, Z_{012}^2 must be negative there. Consequently, if $Z_{I_1}Z_{I_2}$ changes sign, then $\dfrac{1}{\tanh^2 \theta} - 1$ must change sign at the same time, and conversely.

In the pass band, the image impedances are resistances, and the transfer constant is imaginary, and so the requirement on Z_{012}^2 is satisfied. In the attenuating band where the image impedances are reactances, and $\tanh\theta$ is real, there are two possibilities:

(a) If the image impedances are reactances of the same sign, the product $Z_{I_1}Z_{I_2}$ is negative. Therefore, $\tanh\theta$ must be equal to or less than unity. This interval will be called a type I attenuating band and will be designated ABI.

(b) If the image impedances are reactances of opposite sign, $Z_{I_1}Z_{I_2}$ is positive. Therefore, $\tanh\theta$ must be equal to or greater than unity. This interval will be called a type II attenuating band and will be designated ABII.

20. TRANSFER CONSTANT THEOREM

The transfer constant of any physically realizable filter is uniquely determined by the cutoff frequencies and by the frequencies of infinite attenuation.

The frequencies at which the attenuation is infinite are the roots of the equation $\tanh\theta = 1$. In practice, they are found by determining the roots of the equation $\tanh^2\theta - 1 = 0$. The number of roots in terms of ω^2 is equal to the number of pole-zero intervals of $\tanh\theta$ in the pass band or, equivalently, the total phase shift in radians divided by $\pi/2$. From a consideration of the properties of the open-circuit transfer impedance, it is deduced that the admissible roots of a physically realizable filter are the following: (a) roots of even multiplicity located at positive values of ω^2; (b) roots of even multiplicity located at negative values of ω^2; (c) roots of even multiplicity located at conjugate complex values of ω^2; (d) roots of odd multiplicity located at positive values of ω^2. As the later discussion shows, filter sections having roots of even multiplicity are the rule rather than the exception. They are symmetrical sections, for which the restrictions on the image impedance are the minimum. They are usually designed to have the simplest image impedances.

There are two additional restrictions on the expression for $\tanh\theta$: (a) If zero or infinite frequency lies in a pass band, $\tanh\theta$ must have a zero at this frequency; (b) if zero or infinite frequency lies in an attenuating band, $\tanh\theta$ must be equal to or less than unity at this frequency; in particular, $\tanh\theta = 1$, if this frequency lies in an ABII.

The significance of the above theorem is that if an expression for $\tanh\theta$ is found which contains the chosen cutoff frequencies, and is equal to unity at the chosen (admissible) frequencies of infinite attenuation, it is the only such expression that does exist, and it will lead to a physically realizable filter. Such an expression can always be found.

Illustration. (a) Test the expression

$$\tanh\theta = \frac{j\omega\sqrt{2-\omega^2}}{1-\omega^2}$$

for physical realizability. The equation $\tanh\theta = 1$ has two roots. By forming

$$\tanh^2\theta - 1 = \frac{-1}{(1-\omega^2)^2} = 0$$

both are found to be located at $\omega^2 = \infty$. Therefore, this expression has a physical representation.

(b) Test the expression

$$\tanh\theta = \frac{j\omega\sqrt{3-\omega^2}}{1-\omega^2}$$

for physical realizability. From

$$\tanh^2\theta - 1 = \frac{-(1+\omega^2)}{(1-\omega^2)^2} = 0$$

it is seen that there is a single root at $\omega^2 = -1$ and another at $\omega^2 = \infty$. Therefore, this expression for the transfer constant is not realizable.

21. IMAGE IMPEDANCE THEOREM

The second image impedance of a filter is uniquely determined at all frequencies, except for an arbitrary constant multiplier, by the transfer constant and the first image impedance.

This theorem may be demonstrated by an examination of the properties of the open-circuit transfer impedance. The formation of the expression for the second image impedance is accomplished by the application of the following rules:

(a) At the boundary of a pass band and ABI, the cutoff factor appears in the numerator of $\tanh\theta$ and in the numerator or denominator of both Z_{I_1} and Z_{I_2}.

(*b*) At the boundary of a pass band and *AB*II, the cutoff factor appears in the denominator of $\tanh \theta$ and in the numerator of one image impedance and in the denominator of the other image impedance.

(*c*) At the boundary, *AB*I/*AB*II, of the two types of attenuating bands, the equation $\tanh^2 \theta = 1$ has a root of odd multiplicity One image impedance must have a zero, or pole, at this frequency which does not appear in the other image impedance.

(*d*) All other zeros and poles of one image impedance give rise to corresponding zeros and poles (or poles and zeros) in the second image impedance. Attention must be given to the requirement that the zeros and poles of each impedance alternate in an attenuation band.

(*e*) Roots of even multiplicity can be introduced in $(\tanh^2 \theta - 1)$ without regard to the image impedances.

Illustration. The expressions for the transfer constant and the image impedance at one end of a band-pass filter are:

$$\tanh \theta = \sqrt{\frac{\omega_1^2 - \omega^2}{\omega_2^2 - \omega^2}}$$

$$Z_{I_1} = Hj\omega\sqrt{\frac{\omega_2^2 - \omega^2}{\omega_1^2 - \omega^2}}$$

where ω_1 is the lower cutoff, and ω_2 the upper cutoff. The region from zero frequency to ω_1 is an *AB*I, and, by rule (*a*), the cutoff factor corresponding to ω_1 appears in the denominator of Z_{I_2}. The region from ω_2 to infinity is an *AB*II, and, by rule (*b*), the cutoff factor corresponding to ω_2 appears in the denominator of Z_{I_2}. The expression:

$$\tanh^2 \theta - 1 = \frac{\omega_1^2 - \omega_2^2}{\omega_2^2 - \omega^2}$$

has one root in ω^2, and that is at infinity. Thus, there are no other internal zeros or poles in Z_{I_2}. Finally, with the aid of rule (*d*),

$$Z_{I_2} = \frac{Kj\omega}{\sqrt{(\omega_1^2 - \omega^2)(\omega_2^2 - \omega^2)}}$$

The elements of the filter may be found from the short- and open-circuit impedances:

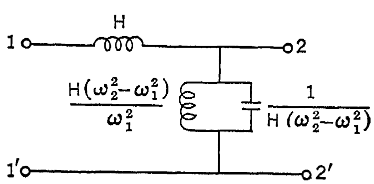


Fig. 4. Illustration for the Method of Obtaining the Second Image Impedance from the Transfer Function and the First Image Impedance

$$Z_{s1} = Z_{I_1} \tanh \theta = Hj\omega$$

$$Z_{01} = \frac{Z_{I_1}}{\tanh \theta} = \frac{Hj\omega(\omega_2^2 - \omega^2)}{(\omega_1^2 - \omega^2)}$$

$$Z_{s2} = Z_{I_2} \tanh \theta = \frac{Kj\omega}{\omega_2^2 - \omega^2}$$

$$Z_{02} = \frac{Z_{I_2}}{\tanh \theta} = \frac{Kj\omega}{\omega_1^2 - \omega^2}$$

The network obtained by setting $K = H(\omega_2^2 - \omega_1^2)$ is shown in Fig. 4.

22. THE GENERAL COMPOSITE FILTER

As exemplified by the preceding illustration, a possible filter design method consists of setting up physically realizable expressions for the desired transfer constant and image impedance at one end of the filter. Then, the expression for the second image impedance is found with the aid of the rules associated with the image impedance theorem. The corresponding short- and open-circuit impedances are computed. From these, it is possible to find the elements of the filter. In general, this may require some ingenuity.

The design method which is used most frequently is based on the fact that every filter can be regarded as the combination of certain elementary sections. As shown in Fig. 5, the composite filter is made up of N elementary sections connected in tandem. Section *A* provides one or more frequencies of infinite attenuation (in terms of ω^2) and has the image impedance Z_{I_1} specified for the input side of the filter. The secondimage im pedance Z_{I_a} is determined by Z_{I_1} and the roots of $(\tanh \theta = 1)$ provided by section *A*. Section *B* has

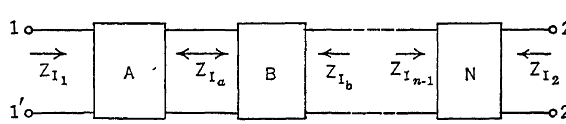


Fig. 5. The General Composite Filter

the image impedance Z_{I_a} to match A and provides one or more of the remaining frequencies of infinite attenuation. Its other image impedance is Z_{I_b}. The last section has the image impedance Z_{I_2}, consistent with the specification of Z_{I_1} and all the frequencies of infinite attenuation, that is, all the roots of $\tanh \theta = 1$.

The first step in the application of this design method is the selection of the terminating sections to furnish the desired image impedance. They provide a simple image impedance (i.e., one without impedance controlling factors) at their inner pairs of terminals. These sections correspond to A and N in Fig. 5. The transfer constant contributed by these sections is subtracted from the required overall transfer constant. The balance is supplied by elementary sections of known transfer constant characteristics which have simple image impedances. These are inserted between the terminating sections.

An alternative design procedure is applicable where it is convenient to establish a physically realizable expression for the transfer function, $\tanh \theta$, which meets the requirements of the design objective. Such a case arises, for example, if the objective is to attain linearity of phase over the pass band; another objective which can be handled in this way, as will be described later in this article, is to provide a prescribed minimum of attenuation over a specified interval of the attenuating band. The next step in the design is the determination of the frequencies of infinite attenuation by solving the equation $\tanh \theta = 1$. Then these frequencies are assigned to the appropriate elementary sections, and the sections are assembled to form a composite filter.

These design procedures are facilitated if there is available a list of the characteristics and element values of the elementary sections and the terminating sections. Such a tabulation should contain sections which provide double positive and negative roots in ω^2, double pairs of conjugate complex roots, simple positive roots. In general, the double roots correspond to the elementary symmetrical sections which form the main body of the filter, and the simple roots correspond to the unsymmetrical sections which are used for terminations. The tabulations of these sections are discussed in the following two paragraphs.

23. SYMMETRICAL SECTIONS

The problem associated with the design of symmetrical filters is much simpler than the general design problem. The configuration which is convenient for analysis is the lattice, shown in Fig. 6. From the fundamental eqs. (1)–(3) for the image parameters, it may be shown that:

$$Z_I = \sqrt{Z_x Z_y} \tag{10}$$

$$\tanh \frac{\theta}{2} = \sqrt{\frac{Z_x}{Z_y}} \tag{11}$$

By identifying Z_x with Z_{s1}, and Z_y with Z_{01}, the analysis in terms of critical frequencies, given in article 18, is directly applicable. The restrictions on Z_I and $\tanh \theta/2$ are simply those summarized there for Z_I and $\tanh \theta$. The only distinction is that the resulting transfer constant, θ, of the lattice is twice the value $\theta/2$ appearing in eq. (11). The restrictions contained in the transfer constant and image impedance theorems are satisfied since all the roots of ($\tanh \theta = 1$) are of even multiplicity. This means that the expressions for the transfer constant and image impedance can be chosen independently except for the cutoff frequencies which are the same in the two expressions. The branch impedances are:

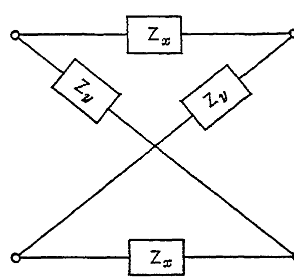


FIG. 6. The General Symmetrical Lattice Configuration

$$Z_x = Z_I \tanh \frac{\theta}{2} \tag{12}$$

$$Z_y = \frac{Z_I}{\tanh (\theta/2)} \tag{13}$$

These impedances satisfy the requirements of Foster's theorem and may be developed into the structures listed in Fig. 2.

It is evident that these branches will contain a large number of elements for all but the simplest filters. Since attenuation is obtained by bridge balance, these elements must be held to close limits if the required attenuation is great. Consequently, though the lattice is much used in theoretical work, it is usually converted, when possible, to other configurations. The first step in the conversion process is to apply the concept of the

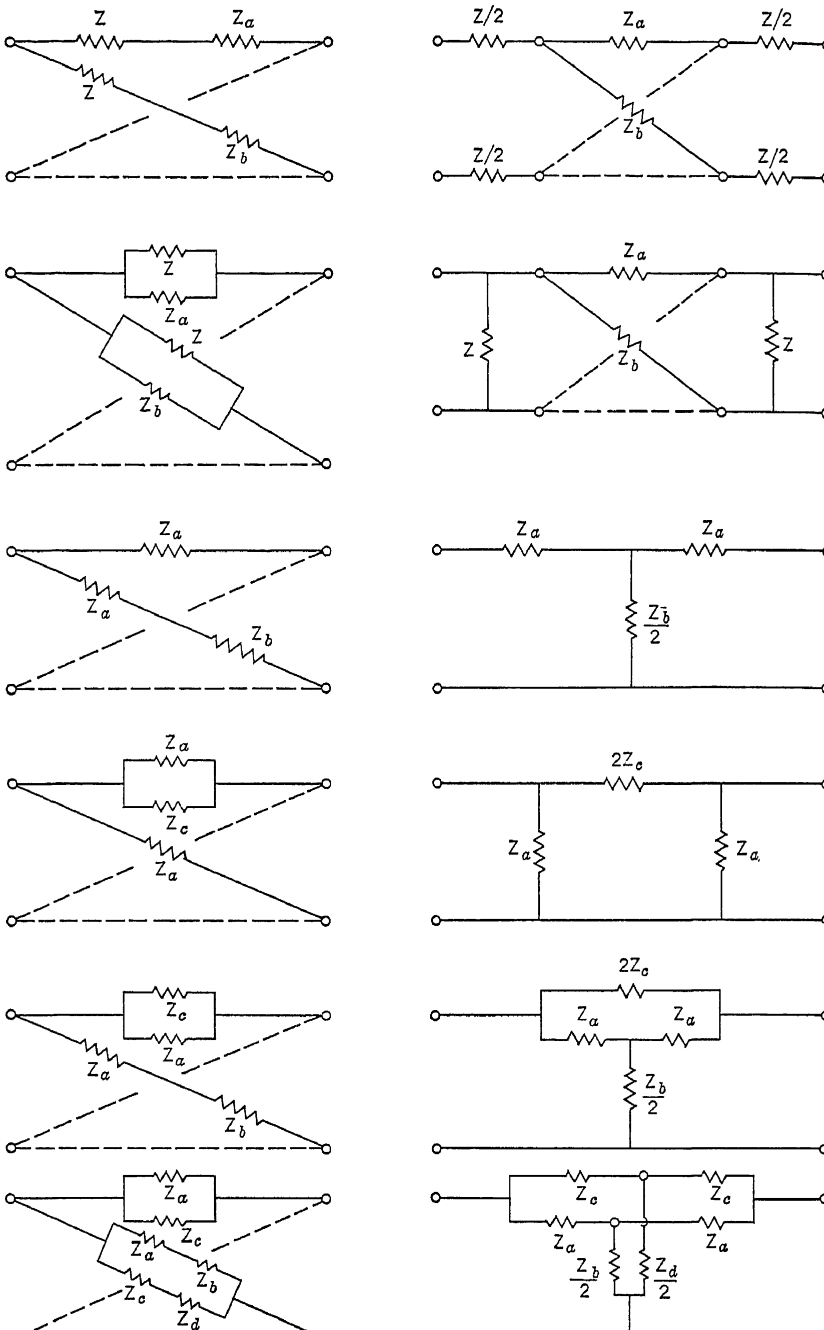


FIG. 7. Conversions of the Symmetrical Lattice

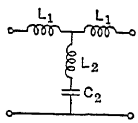
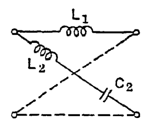
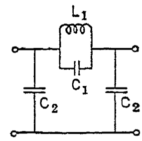
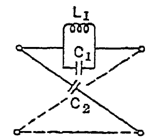
	(1)	(2)	(3)	(4)
				
L_1	$\frac{mR}{2\pi f_c}$	$\frac{mR}{2\pi f_c}$	$\frac{mR}{\pi f_c}$	$\frac{mR}{2\pi f_c}$
L_2	$\frac{(1-m^2)R}{4m\pi f_c}$	$\frac{R}{2m\pi f_c}$	....	....
C_1	....	....	$\frac{(1-m^2)}{4m\pi f_c R}$	$\frac{1}{2m\pi f_c R}$
C_2	$\frac{m}{\pi f_c R}$	$\frac{m}{2\pi f_c R}$	$\frac{m}{2\pi f_c R}$	$\frac{m}{2\pi f_c R}$
m	$\sqrt{1-\frac{f_c^2}{f_\infty^2}}$ $0 < m \leq 1$	$\sqrt{1-\frac{f_c^2}{f_\infty^2}}$ $0 < m < \infty$	$\sqrt{1-\frac{f_c^2}{f_\infty^2}}$ $0 < m \leq 1$	$\sqrt{1-\frac{f_c^2}{f_\infty^2}}$ $0 < m < \infty$
Image impedance Z_I	$\frac{R}{f_c}\sqrt{f_c^2-f^2}$	Same as (1)	$\frac{Rf_c}{\sqrt{f_c^2-f^2}}$	Same as (3)
$\tanh\frac{\theta}{2}$	$\frac{jmf}{\sqrt{f_c^2-f^2}}$	Same as (1)	Same as (1)	Same as (1)
Special case $m = 1$ $f_\infty = \infty$	L_2 = short circuit	$L_1 = L_2$	C_1 = open circuit	$C_1 = C_2$

FIG. 8. Design Information for Elementary Symmetrical Low-pass Filter Sections

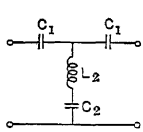
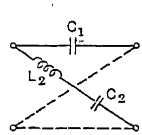
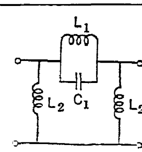
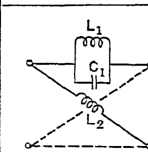
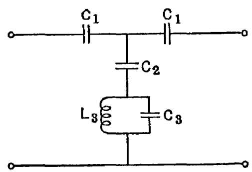
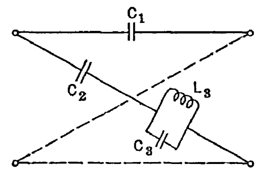
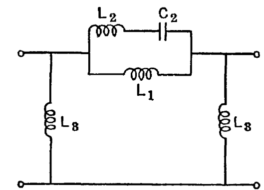
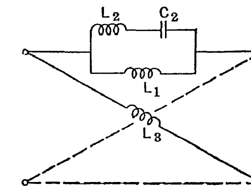
	(1)	(2)	(3)	(4)
				
L_1	....	....	$\frac{mR}{(1-m^2)\pi f_c}$	$\frac{mR}{2\pi f_c}$
L_2	$\frac{R}{4m\pi f_c}$	$\frac{R}{2m\pi f_c}$	$\frac{R}{2m\pi f_c}$	$\frac{R}{2m\pi f_c}$
C_1	$\frac{1}{2m\pi f_c R}$	$\frac{1}{2m\pi f_c R}$	$\frac{1}{4m\pi f_c R}$	$\frac{1}{2m\pi f_c R}$
C_2	$\frac{m}{(1-m^2)\pi f_c R}$	$\frac{m}{2\pi f_c R}$	....	....
m	$\sqrt{1-\frac{f_\infty^2}{f_c^2}}$ $0 < m \leq 1$	$\sqrt{1-\frac{f_\infty^2}{f_c^2}}$ $0 < m < \infty$	$\sqrt{1-\frac{f_\infty^2}{f_c^2}}$ $0 < m \leq 1$	$\sqrt{1-\frac{f_\infty^2}{f_c^2}}$ $0 < m < \infty$
Image impedance Z_I	$\frac{R}{jf}\sqrt{f_c^2-f^2}$	Same as (1)	$\frac{jfR}{\sqrt{f_c^2-f^2}}$	Same as (3)
$\tanh\frac{\theta}{2}$	$\frac{mf_c}{\sqrt{f_c^2-f^2}}$	Same as (1)	Same as (1)	Same as (1)
Special case $m = 1$ $f_\infty = 0$	C_2 = short circuit	$C_1 = C_2$	L_1 = open circuit	$L_1 = L_2$

FIG. 9. Design Information for Elementary Symmetrical High-pass Filter Sections

	(1)	(2)	(3)	(4)
				
L_1	$\frac{mf_1R}{2\pi f_2(f_2 - f_1)}$	$\frac{mf_1R}{2\pi f_2(f_2 - f_1)}$	$\frac{m(f_2 - f_1)R}{(1 - m^2)\pi f_1 f_2}$	$\frac{m(f_2 - f_1)R}{2\pi f_1 f_2}$
L_2	$\frac{(f_2^2 - m^2 f_1^2)R}{4m\pi f_1 f_2(f_2 - f_1)}$	$\frac{f_2R}{2m\pi f_1(f_2 - f_1)}$	$\frac{(f_2 - f_1)R}{2m\pi f_1 f_2}$	$\frac{(f_2 - f_1)R}{2m\pi f_1 f_2}$
L_3	....	....	....	....
C_1	$\frac{(f_2 - f_1)}{2m\pi f_1 f_2 R}$	$\frac{(f_2 - f_1)}{2m\pi f_1 f_2 R}$	$\frac{(f_2^2 - m^2 f_1^2)}{4m\pi f_1 f_2(f_2 - f_1)R}$	$\frac{f_2}{2m\pi f_1(f_2 - f_1)R}$
C_2	$\frac{m(f_2 - f_1)}{(1 - m^2)\pi f_1 f_2 R}$	$\frac{m(f_2 - f_1)}{2\pi f_1 f_2 R}$	$\frac{mf_1}{2\pi f_2(f_2 - f_1)R}$	$\frac{mf_1}{2\pi f_2(f_2 - f_1)R}$
C_3	....	....	....	....
m	$\frac{f_2}{f_1}\sqrt{\frac{f_1^2 - f_\infty^2}{f_2^2 - f_\infty^2}}$ $0 < m \leq 1$	Same as (1) $0 < m < \infty$	Same as (1) $0 < m \leq 1$	Same as (1) $0 < m < \infty$
Image impedance Z_I	$\frac{R\sqrt{(f_1^2 - f^2)(f_2^2 - f^2)}}{jf(f_2 - f_1)}$	Same as (1)	$\frac{jf(f_2 - f_1)R}{\sqrt{(f_1^2 - f^2)(f_2^2 - f^2)}}$	Same as (3)
$\tanh \frac{\theta}{2}$	$m\frac{f_1}{f_2}\sqrt{\frac{f_2^2 - f^2}{f_1^2 - f^2}}$	Same as (1)	Same as (1)	Same as (1)
Special case $m = 1$ $f_\infty = 0$	C_2 = short circuit	$C_1 = C_2$	L_1 = open circuit	$L_1 = L_2$

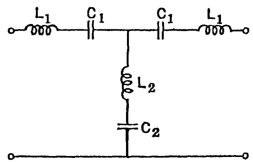
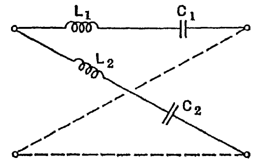
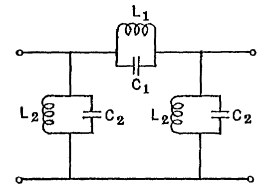
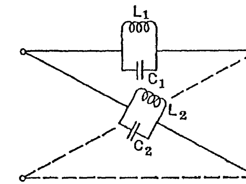
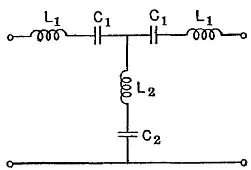
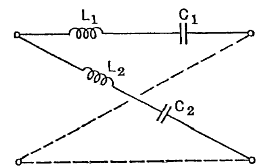
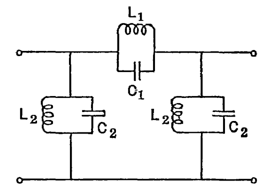
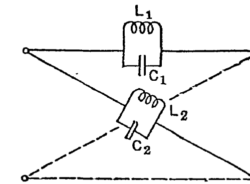
	(5)	(6)	(7)	(8)
				
L_1	....	....	$\dfrac{mR}{(1-m^2)\pi f_1}$	$\dfrac{mR}{2\pi f_1}$
L_2	....	....	$\dfrac{mf_1R}{\pi(f_2^2-f_1^2)}$	$\dfrac{mf_1R}{2\pi(f_2^2-f_1^2)}$
L_3	$\dfrac{(f_2^2-f_1^2)R}{4m\pi f_1f_2^2}$	$\dfrac{(f_2^2-f_1^2)R}{2m\pi f_1f_2^2}$	$\dfrac{R}{2m\pi f_1}$	$\dfrac{R}{2m\pi f_1}$
C_1	$\dfrac{1}{2m\pi f_1R}$	$\dfrac{1}{2m\pi f_1R}$	....	....
C_2	$\dfrac{m}{(1-m^2)\pi f_1R}$	$\dfrac{m}{2\pi f_1R}$	$\dfrac{(f_2^2-f_1^2)}{4m\pi f_1f_2^2R}$	$\dfrac{(f_2^2-f_1^2)}{2m\pi f_1f_2^2R}$
C_3	$\dfrac{mf_1}{\pi(f_2^2-f_1^2)R}$	$\dfrac{mf_1}{2\pi(f_2^2-f_1^2)R}$	....	....
m	Same as (1) $0 < m \leq 1$	Same as (1) $0 < m < \infty$	Same as (1) $0 < m \leq 1$	Same as (1) $0 < m < \infty$
Image impedance Z_I	$\dfrac{f_2R}{jf}\sqrt{\dfrac{f_1^2-f^2}{f_2^2-f^2}}$	Same as (5)	$\dfrac{jfR}{f_2}\sqrt{\dfrac{f_2^2-f^2}{f_1^2-f^2}}$	Same as (7)
$\tanh\frac{\theta}{2}$	Same as (1)	Same as (1)	Same as (1)	Same as (1)
Special case $m = 1$, $f_\infty = 0$	C_2 = short circuit	$C_1 = C_2$	L_1 = open circuit	$L_1 = L_3$

FIG. 10a. Design Information for Elementary Symmetrical Band-pass Filter Sections. (Attenuation peaks of ladder-type sections are located below the lower cutoff frequency.)

	(1)	(2)	(3)	(4)
				
L_1	$\frac{f_2 R}{2m\pi f_1(f_2 - f_1)}$	$\frac{f_2 R}{2m\pi f_1(f_2 - f_1)}$	$\frac{m(f_2 - f_1)R}{(m^2 - 1)\pi f_1 f_2}$	$\frac{(f_2 - f_1)R}{2m\pi f_1 f_2}$
L_2	$\frac{(f_2{}^2 - m^2 f_1{}^2)R}{4m\pi f_1 f_2(f_2 - f_1)}$	$\frac{mf_1 R}{2\pi f_2(f_2 - f_1)}$	$\frac{m(f_2 - f_1)R}{2\pi f_1 f_2}$	$\frac{m(f_2 - f_1)R}{2\pi f_1 f_2}$
L_3	....	....	....	....
C_1	$\frac{m(f_2 - f_1)}{2\pi f_1 f_2 R}$	$\frac{m(f_2 - f_1)}{2\pi f_1 f_2 R}$	$\frac{(m^2 f_1{}^2 - f_2{}^2)}{4m\pi f_1 f_2(f_2 - f_1)R}$	$\frac{mf_1}{2\pi f_2(f_2 - f_1)R}$
C_2	$\frac{m(f_2 - f_1)}{(m^2 - 1)\pi f_1 f_2 R}$	$\frac{(f_2 - f_1)}{2m\pi f_1 f_2 R}$	$\frac{f_2}{2m\pi f_1(f_2 - f_1)R}$	$\frac{f_2}{2m\pi f_1(f_2 - f_1)R}$
C_3	....	....	....	....
m	$\frac{f_2}{f_1}\sqrt{\frac{f_1{}^2 - f_\infty{}^2}{f_2{}^2 - f_\infty{}^2}}$ $\frac{f_2}{f_1} \leq m < \infty$	Same as (1) $0 < m < \infty$	Same as (1) $\frac{f_2}{f_1} \leq m < \infty$	Same as (1) $0 < m < \infty$
Image impedance Z_I	$\frac{R\sqrt{(f_1{}^2 - f^2)(f_2{}^2 - f^2)}}{jf(f_2 - f_1)}$	Same as (1)	$\frac{jf(f_2 - f_1)R}{\sqrt{(f_1{}^2 - f^2)(f_2{}^2 - f^2)}}$	Same as (3)
$\tanh\frac{\theta}{2}$	$\frac{1}{m}\frac{f_2}{f_1}\sqrt{\frac{f_1{}^2 - f^2}{f_2{}^2 - f^2}}$	Same as (1)	Same as (1)	Same as (1)
Special case $m = \frac{f_2}{f_1}$ $f_\infty = \infty$	L_2 = Short circuit	$L_1 = L_2$	C_1 = open circuit	$C_1 = C_2$

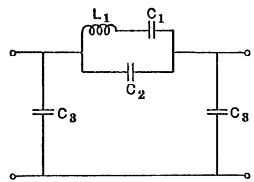
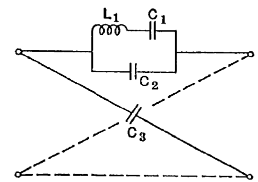
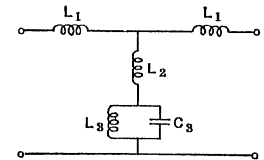
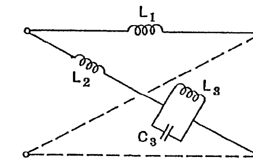
	(5)	(6)	(7)	(8)
				
L_1	$\dfrac{f_2^2 R}{m\pi f_1(f_2^2 - f_1^2)}$	$\dfrac{f_2^2 R}{2m\pi f_1(f_2^2 - f_1^2)}$	$\dfrac{R}{2m\pi f_1}$	$\dfrac{R}{2m\pi f_1}$
L_2	....	....	$\dfrac{(m^2 f_1^2 - f_2^2)R}{4m\pi f_1 f_2^2}$	$\dfrac{m f_1 R}{2\pi f_2^2}$
L_3	....	....	$\dfrac{m(f_2^2 - f_1^2)R}{4\pi f_1 f_2^2}$	$\dfrac{m(f_2^2 - f_1^2)R}{2\pi f_1 f_2^2}$
C_1	$\dfrac{m(f_2^2 - f_1^2)}{4\pi f_1 f_2^2 R}$	$\dfrac{m(f_2^2 - f_1^2)}{2\pi f_1 f_2^2 R}$	....	....
C_2	$\dfrac{(m^2 f_1^2 - f_2^2)}{4m\pi f_1 f_2^2 R}$	$\dfrac{m f_1}{2\pi f_2^2 R}$	....	....
C_3	$\dfrac{1}{2m\pi f_1 R}$	$\dfrac{1}{2m\pi f_1 R}$	$\dfrac{f_2^2}{m\pi f_1(f_2^2 - f_1^2)R}$	$\dfrac{f_2^2}{2m\pi f_1(f_2^2 - f_1^2)R}$
m	Same as (1) $\dfrac{f_2}{f_1} \leq m < \infty$	Same as (1) $0 < m < \infty$	Same as (1) $\dfrac{f_2}{f_1} \leq m < \infty$	Same as (1) $0 < m < \infty$
Image impedance Z_I	$\dfrac{f_2 R}{jf}\sqrt{\dfrac{f_1^2 - f^2}{f_2^2 - f^2}}$	Same as (5)	$\dfrac{jfR}{f_2}\sqrt{\dfrac{f_2^2 - f^2}{f_1^2 - f^2}}$	Same as (7)
$\tanh \dfrac{\theta}{2}$	Same as (1)	Same as (1)	Same as (1)	Same as (1)
Special case $m = \dfrac{f_2}{f_1}$ $f_\infty = \infty$	C_2 = open circuit	$C_2 = C_3$	L_2 = short circuit	$L_1 = L_2$

FIG. 10b. Design Information for Elementary Symmetrical Band-pass Filter Sections. (Attenuation peaks of ladder-type sections are located above the upper cutoff frequency.)

composite filter. Then, after the original lattice is separated into a combination of the simplest lattices, these, in turn, are converted to other configurations. Some of the most useful conversions, based on Bartlett's bisection theorem, are given in Fig. 7. (Broken lines are used to simplify the drawings. It is understood that the lattice branches represented by these lines are duplicates of the corresponding ones shown explicitly.)

ELEMENTARY STRUCTURES. The elementary symmetrical sections which provide the double positive and negative roots required in forming a composite filter are listed in Figs. 8, 9, 10. Since a double root of $(\tanh^2 \theta - 1)$ corresponds to a simple root of $\left(\tanh^2 \frac{\theta}{2} - 1\right)$, it is possible to use the more convenient expression, $\tanh(\theta/2)$, in describing these sections. Each of the configurations provides one double peak of infinite attenuation. The associated image impedances are the simplest possible. The transfer constant, $\theta = \alpha + j\beta$, where α is the attenuation constant in nepers, and β is the phase constant in radians, is computed from the expression for $\tanh(\theta/2)$. If the arithmetical value of this expression is denoted by Q, then, in the pass band:

$$\tanh \frac{\theta}{2} = jQ$$

$$\alpha = 0$$

$$\beta = 2 \operatorname{arc} \tan Q; \quad \left(\frac{d\beta}{d\omega} \geq 0\right)$$

In the attenuating band:

$$\tanh \frac{\theta}{2} = Q; \quad (Q \geq 0)$$

There are two possibilities, depending on the value of Q relative to unity. Either:

$$\alpha = 2 \arg \tanh Q; \quad (Q \leq 1)$$

$$\beta = 0$$

or

$$\alpha = 2 \arg \tanh \frac{1}{Q}; \quad (Q \geq 1)$$

$$\beta = \pm\pi$$

The element values specified in these tabulations, as well as all that follow, apply to the filter sections as drawn. That is, each section is considered to be a building block in the composite filter. For example, if two ladder-type mid-series terminated sections having the same image impedance are joined together, the intermediate series impedance becomes equal to the sum of the values given in the figures; similarly, if two ladder-type mid-shunt terminated sections are joined together, the intermediate shunt admittance becomes equal to the sum of the values given.

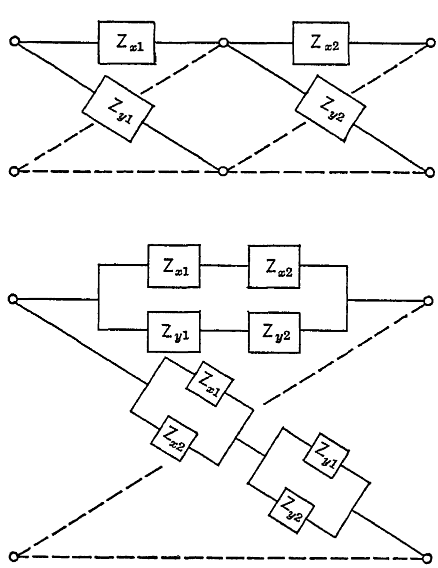


FIG. 11. Combination of Two Lattice Sections Which Have the Same Image Impedance

Figures 8 and 9 contain the design information for low-pass and high-pass sections, respectively. The cutoff frequency is denoted by f_c and the frequency of infinite attenuation by f_∞. The image impedance is equal to R at zero frequency for the low pass, and at infinite frequency for the high pass. The sections numbered (1) and (3) provide double peaks of attenuation at real frequencies. They are the m-derived sections introduced by O. J. Zobel. The special case for which $m = 1$ is the constant-K section. Sections (2) and (4) provide double peaks at real frequencies for values of the parameter m lying in the range $0 < m \leq 1$, and double peaks at imaginary values of frequency (i.e., negative values of ω^2) for values of m greater than unity.

The elementary band-pass sections are shown in Figs. 10a and b. The lower cutoff frequency is denoted by f_1, the upper cutoff by f_2, and the peak of infinite attenuation by f_∞. The image impedance is equal to R at the mid-band frequency, $f_m = \sqrt{f_1 f_2}$. A

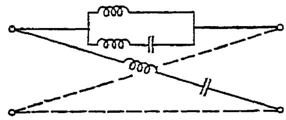
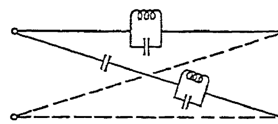
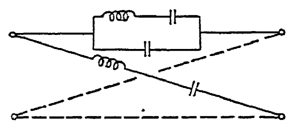
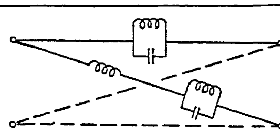
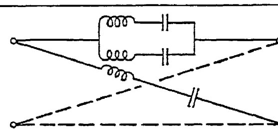
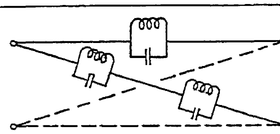
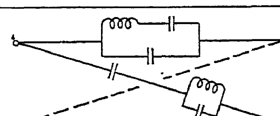
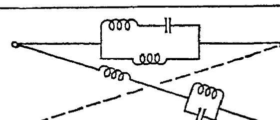
Type	Image Impedance	Configuration
Low-pass	$\frac{R}{f_c}\sqrt{f_c^2 - f^2}$	
	$\frac{Rf_c}{\sqrt{f_c^2 - f^2}}$	
High-pass	$\frac{R}{jf}\sqrt{f_c^2 - f^2}$	
	$\frac{jfR}{\sqrt{f_c^2 - f^2}}$	
Band-pass	$\frac{R\sqrt{(f_1^2 - f^2)(f_2^2 - f^2)}}{jf(f_2 - f_1)}$	
	$\frac{jf(f_2 - f_1)R}{\sqrt{(f_1^2 - f^2)(f_2^2 - f^2)}}$	
	$\frac{f_2 R}{jf}\sqrt{\frac{f_1^2 - f^2}{f_2^2 - f^2}}$	
	$\frac{jfR}{f_2}\sqrt{\frac{f_2^2 - f^2}{f_1^2 - f^2}}$	

FIG. 12. Elementary Constituents of the General Composite Filter Which Provide Attenuation Peaks at Complex Values of Frequency

uniform definition for the parameter m is used throughout, with the result that the range of values is extended beyond the conventional zero to unity. The odd-numbered sections in Fig. 10a correspond to the usual m-derived sections which have a peak of attenuation below the lower cutoff. The special cases, for which $m = 1$, are the so-called three-element type band-pass sections having a peak at zero frequency. The lattice sections provide a peak below the lower cutoff for $0 < m \leq 1$, and a peak above the upper cutoff for $f_2/f_1 \leq m < \infty$. For the range $1 < m < f_2/f_1$, the peak is located at an imaginary value of frequency. In all cases, the phase shift is $(-\pi)$ radians at f_1 and zero at f_2.

The odd-numbered sections in Fig. 10*b* correspond to the usual m-derived sections which have a peak of attenuation above the upper cutoff. The three-element type sections having a peak at infinite frequency are special cases, for which $m = f_2/f_1$. The lattice sections are the same as those shown in Fig. 10*a* except that the branches are interchanged. In all cases, the phase shift is zero at f_1, and $(+\pi)$ radians at f_2.

The elementary sections which provide the double pairs of conjugate complex roots have a double peak of attenuation at the complex value of frequency, f_∞, and a double peak at the conjugate value, $\bar{f}_\infty$. They can be derived by combining two lattice sections of the type tabulated in Figs. 8–10. The method for combining two symmetrical lattice sections which have the same image impedance is shown in Fig. 11. By using the definitions for the parameter m given in the previous figures, a complex value, $m = m_1 + jm_2$, is obtained for the section which provides the peak at the complex value of frequency, f_∞; the conjugate complex value, $\bar{m} = m_1 - jm_2$, is obtained for the section which provides the peak at the conjugate complex value, $\bar{f}_\infty$. (The real part, m_1, must be positive.) The elements of the individual lattice sections are proportional to m and $\bar{m}$, and their reciprocals, and therefore are complex quantities. However, upon combining the two sections, the resulting elements are ordinary coils and condensers. These sections are displayed in Fig. 12.

24. UNSYMMETRICAL SECTIONS

The unsymmetrical sections provide the simple positive roots of $(\tanh^2 \theta - 1)$ in terms of ω^2. Generally, they are viewed as the means for converting the simple image impedance of the main part of the composite filter into an image impedance which approximates as closely as required to a constant value, equal to the resistance in which the filter is terminated. For low- and high-pass filters, the conversion is from a constant-K type image impedance to one having one or more impedance-controlling frequencies. For band-pass filters, there is a greater variety of possibilities. However, as a practical matter, the terminating sections are usually designed to convert a constant-K image impedance into one having one or more pairs of impedance-controlling frequencies. The product of the frequencies making up a pair is equal to the square of the mid-band frequency. For completeness, it is necessary to have available the simple sections which convert a three-element type image impedance into the geometrically symmetrical constant-K type.

The design information for the simple sections which are used to obtain either a mid-series or mid-shunt constant-K image impedance is given in Fig. 13. They are all "half-sections," and the element values apply to the sections as drawn. Comparison of the formulas for the element values and image impedance of the three-element band-pass sections with those given for the special cases in Figs. 10*a*–*b* shows that they differ by a factor which is a function of the cutoff frequencies. This arises from the fact that the symmetrical sections are designed to give an image impedance equal to R at the mid-band frequency, while those in Fig. 13 satisfy this condition only at the constant-K end of the structure. Hence, the impedance level of one or the other must be changed by the factor specified in the lower part of Fig. 13 before they can be joined without reflection. In general, the level of the symmetrical sections is changed so that the constant-K impedance of the terminating section has a mid-band value equal to the termination. For example, the inductances of sections 7 and 8 of Fig. 10*a* are multiplied by the factor $(f_2 + f_1)/f_2$, and the capacitances divided by this factor, if the section is joined to the first band-pass section of Fig. 13.

***m*-DERIVED SECTIONS.** The design information for terminating sections which present a constant-K image impedance at one end and an image impedance having one controlling frequency (or one geometrically symmetrical pair) at the other end is given in Fig. 14. It follows from the rules associated with the image impedance theorem that the low- and high-pass sections have one simple attenuation peak at the impedance controlling frequency, and the band-pass sections have a geometrically symmetrical pair. It is convenient to use a universal frequency variable, denoted by x, in the description of all the sections. The definition of this variable for each type of filter is included in the figure. For the low-pass sections, the pass band extends from $x = 0$ to $+1$, and the attenuation band from $x = +1$ to plus infinity. For the high-pass sections, the pass band extends from $x = 0$ to $x = -1$, and the attenuation band from $x = -1$ to minus infinity. For the band-pass sections, the pass band extends from $x = -1$ to $x = +1$ with the mid-band at $x = 0$; the attenuation band above the upper cutoff extends from $x = +1$ to plus infinity, while the attenuation band below the lower cutoff extends from $x = -1$ to minus infinity.

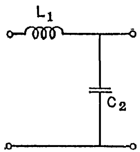
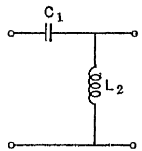
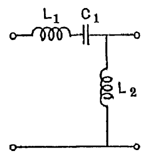
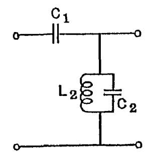
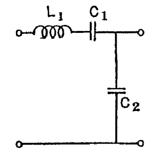
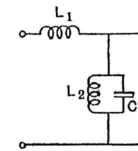
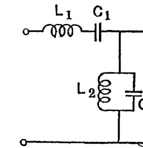
Type	Low-pass Constant-K	High-pass Constant-K	Band-pass 3-element	Band-pass 3-element	Band-pass 3-element	Band-pass 3-element	Band-pass Constant-K
Configuration							
L_1	$\frac{R}{2\pi f_c}$	...	$\frac{f_1R}{2\pi f_2(f_2-f_1)}$	...	$\frac{R}{2\pi(f_2-f_1)}$	$\frac{R}{2\pi(f_2+f_1)}$	$\frac{R}{2\pi(f_2-f_1)}$
L_2	...	$\frac{R}{2\pi f_c}$	$\frac{(f_2+f_1)R}{2\pi f_1f_2}$	$\frac{(f_2-f_1)R}{2\pi f_1f_2}$	...	$\frac{(f_2-f_1)R}{2\pi f_1^2}$	$\frac{(f_2-f_1)R}{2\pi f_1f_2}$
C_1	...	$\frac{1}{2\pi f_cR}$	$\frac{(f_2-f_1)}{2\pi f_1f_2R}$	$\frac{(f_2+f_1)}{2\pi f_1f_2R}$	$\frac{(f_2-f_1)}{2\pi f_1^2R}$	...	$\frac{(f_2-f_1)}{2\pi f_1f_2R}$
C_2	$\frac{1}{2\pi f_cR}$	...	...	$\frac{f_1}{2\pi f_2(f_2-f_1)R}$	$\frac{1}{2\pi(f_2+f_1)R}$	$\frac{1}{2\pi(f_2-f_1)R}$	$\frac{1}{2\pi(f_2-f_1)R}$
Mid-series image impedance Z_I	$\frac{R}{f_c}\sqrt{f_c^2-f^2}$	$\frac{R}{jf}\sqrt{f_c^2-f^2}$	$\frac{R\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}{jf(f_2-f_1)}$	$\frac{f_2^2R}{jf(f_2+f_1)}\sqrt{\frac{f_1^2-f^2}{f_2^2-f^2}}$	$\frac{R\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}{jf(f_2-f_1)}$	$\frac{jfR}{(f_2+f_1)}\sqrt{\frac{f_2^2-f^2}{f_1^2-f^2}}$	$\frac{R\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}{jf(f_2-f_1)}$
Mid-shunt image impedance Z_I'	$\frac{Rf_c}{\sqrt{f_c^2-f^2}}$	$\frac{jfR}{\sqrt{f_c^2-f^2}}$	$\frac{jf(f_2+f_1)R}{f_2^2}\sqrt{\frac{f_2^2-f^2}{f_1^2-f^2}}$	$\frac{jf(f_2-f_1)R}{\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}$	$\frac{(f_2+f_1)R}{jf}\sqrt{\frac{f_1^2-f^2}{f_2^2-f^2}}$	$\frac{jf(f_2-f_1)R}{\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}$	$\frac{jf(f_2-f_1)R}{\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}$
$\tanh\theta$	$\frac{jf}{\sqrt{f_c^2-f^2}}$	$\frac{f_c}{\sqrt{f_c^2-f^2}}$	$\frac{f_1}{f_2}\sqrt{\frac{f_2^2-f^2}{f_1^2-f^2}}$	$\frac{f_1}{f_2}\sqrt{\frac{f_2^2-f^2}{f_1^2-f^2}}$	$\sqrt{\frac{f_1^2-f^2}{f_2^2-f^2}}$	$\sqrt{\frac{f_1^2-f^2}{f_2^2-f^2}}$	$\frac{(f_1f_2-f^2)}{\sqrt{(f_1^2-f^2)(f_2^2-f^2)}}$
Z_I at $f_m=\sqrt{f_1f_2}$	...	...	R	$\frac{f_2}{(f_2+f_1)}\cdot R$	R	$\frac{f_2}{(f_2+f_1)}\cdot R$	R
Z_I' at $f_m=\sqrt{f_1f_2}$	...	...	$\frac{(f_2+f_1)}{f_2}\cdot R$	R	$\frac{(f_2+f_1)}{f_2}\cdot R$	R	R

FIG. 13. Design Information for Simple Sections Used to Obtain Either a Mid-series or Mid-shunt Constant-K Image Impedance

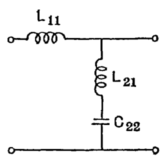
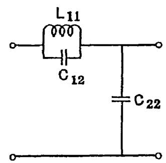
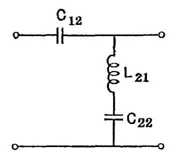
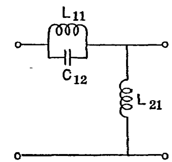
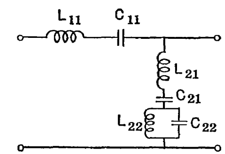
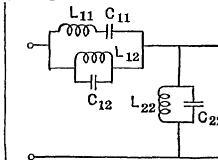
Type	(1)	(2)	(3)	(4)	(5)	(6)
	Low-pass		High-pass		Band-pass	
Configuration						
L_{11}	$\frac{mR}{2\pi f_c}$	$\frac{mR}{2\pi f_c}$	....	$\frac{mR}{2(1-m^2)\pi f_c}$	$\frac{mR}{2\pi(f_2-f_1)} = \frac{1}{4\pi^2 f_m^2 C_{11}}$	$\frac{mR}{2\pi(f_2-f_1)} = \frac{1}{4\pi^2 f_m^2 C_{11}}$
L_{21}	$\frac{(1-m^2)R}{2m\pi f_c}$	....	$\frac{R}{2m\pi f_c}$	$\frac{R}{2m\pi f_c}$	$\frac{(1-m^2)R}{2m\pi(f_2-f_1)} = \frac{1}{4\pi^2 f_m^2 C_{21}}$	....
C_{12}	....	$\frac{(1-m^2)}{2m\pi f_c R}$	$\frac{1}{2m\pi f_c R}$	$\frac{1}{2m\pi f_c R}$	....	$\frac{(1-m^2)}{2m\pi(f_2-f_1)R} = \frac{1}{4\pi^2 f_m^2 L_{12}}$
C_{22}	$\frac{m}{2\pi f_c R}$	$\frac{m}{2\pi f_c R}$	$\frac{m}{2(1-m^2)\pi f_c R}$	....	$\frac{m}{2\pi(f_2-f_1)R} = \frac{1}{4\pi^2 f_m^2 L_{22}}$	$\frac{m}{2\pi(f_2-f_1)R} = \frac{1}{4\pi^2 f_m^2 L_{22}}$
m	$\sqrt{1-\frac{1}{x_\infty^2}}$					
Mid-series image impedance	$R\sqrt{1-x^2}$	$\frac{R\sqrt{1-x^2}}{[1-x^2(1-m^2)]}$	Same as (1)	Same as (2)	Same as (1)	Same as (2)
Mid-shunt image impedance	$\frac{R[1-x^2(1-m^2)]}{\sqrt{1-x^2}}$	$\frac{R}{\sqrt{1-x^2}}$	Same as (1)	Same as (2)	Same as (1)	Same as (2)
$\tanh\theta$	$\frac{jmx}{\sqrt{1-x^2}}$					
x	$\frac{f}{f_c}$		$-\frac{f_c}{f}$		$\frac{f^2-f_m^2}{f(f_2-f_1)}$; $f_m^2 = f_1 f_2$	
x_∞^2	$\frac{f_\infty^2}{f_c^2}$		$\frac{f_c^2}{f_\infty^2}$		$\left[\frac{f_\infty^2-f_m^2}{f_\infty(f_2-f_1)}\right]^2$; $f_\infty > f_m > f_{-\infty}$; $f_\infty f_{-\infty} = f_m^2$	

FIG. 14. Design Information for m-Derived Terminating Sections

The choice of a particular value of the parameter m is dictated by the image impedance characteristic that is desired (on the assumption that the location of the associated attenuation peak is satisfactory). Several characteristics are shown in Fig. 15 including the constant-K type, for which $m = 1.0$. These curves exhibit the course of the mid-series image impedance and the mid-shunt image admittance as a function of the frequency variable, x. A generally satisfactory value of the parameter is $m = 0.6$. It is seen that the image impedance of this section remains within about 4 per cent of a constant value over 87 per cent of the pass band. The actual impedance measured at the terminals of

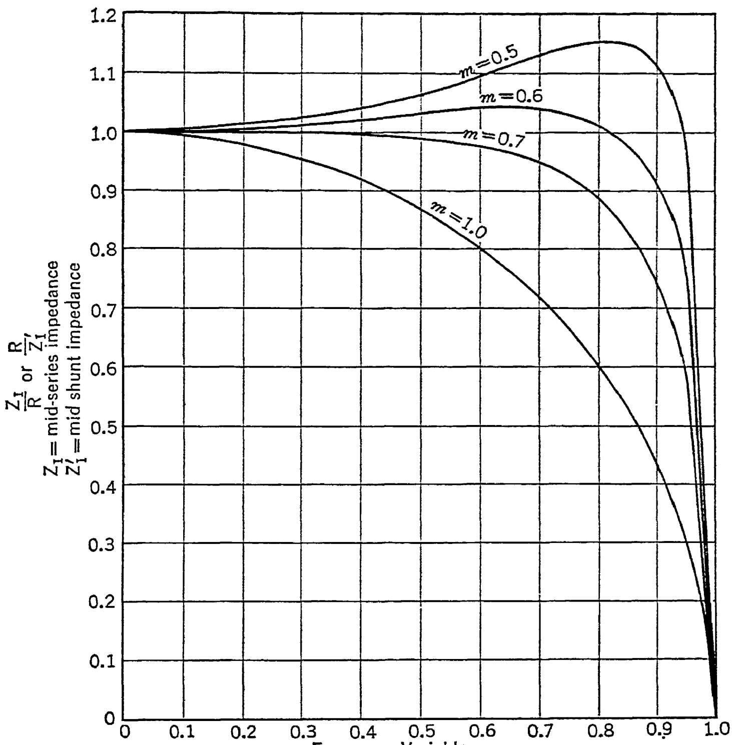


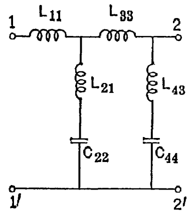
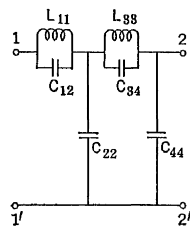
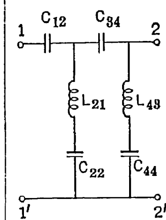
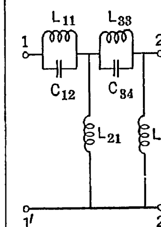
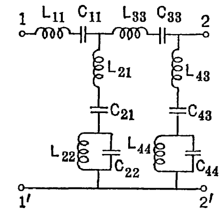
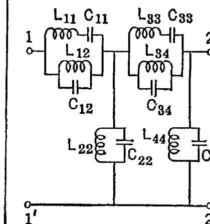
FIG. 15. Image Impedance Characteristics of Several m-derived Filter Sections

the filter is identical with the image impedance only if the other end of the filter is terminated in the image impedance at that end. A general formula for computing this impedance is

$$Z = Z_{I_1}\left[\frac{R_t + Z_{I_2}\tanh\theta}{Z_{I_2} + R_t\tanh\theta}\right]$$

where Z_I is the image impedance at the end of the filter for which the driving point impedance Z is being calculated; Z_{I_2} is the image impedance at the far end which is terminated by the load impedance R_t; and θ is the total image transfer constant of the complete filter. Evidently, if $Z_{I_2} = R_t$, then $Z = Z_{I_1}$.

TWO-FREQUENCY CONTROL SECTIONS. Terminating sections which present an image impedance having two controlling frequencies may be derived in a number of ways and exist in various forms. The sections listed in Fig. 16 serve as a direct transition from a constant-K image impedance to a "two-frequency control" image impedance. They are units consisting of four branches which cannot be separated at any internal junction, even though they have the structural appearance of 1 1/2 section m-derived filters. The low- and high-pass sections have two simple attenuation peaks, and the band-pass have two geometrically symmetrical pairs of simple peaks. The behavior of

Type	(1) Low-pass	(2) Low-pass	(3) High-pass	(4) High-pass	(5) Band-pass	(6) Band-pass
Configuration						
L_{11}	$\dfrac{m_1R}{2\pi f_c}$	$\dfrac{(m_2^2-m_1^2)R}{2m_2\pi f_c}$	...	$\dfrac{(m_2^2-m_1^2)R}{2m_2(1-m_2^2)\pi f_c}$	$\dfrac{m_1R}{2\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{11}}$	$\dfrac{(m_2^2-m_1^2)R}{2m_2\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{11}}$
L_{21}	$\dfrac{m_2(1-m_1^2)R}{2m_1(m_1+m_2)\pi f_c}$	...	$\dfrac{m_2R}{2m_1(m_1+m_2)\pi f_c}$	$\dfrac{R}{2m_2\pi f_c}$	$\dfrac{m_2(1-m_1^2)R}{2m_1(m_1+m_2)\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{21}}$	...
L_{33}	$\dfrac{m_2R}{2\pi f_c}$	$\dfrac{m_1(m_1+m_2)R}{2m_2\pi f_c}$	...	$\dfrac{m_1(m_1+m_2)R}{2m_2(1-m_1^2)\pi f_c}$	$\dfrac{m_2R}{2\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{33}}$	$\dfrac{m_1(m_1+m_2)R}{2m_2\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{33}}$
L_{43}	$\dfrac{m_2(1-m_2^2)R}{2(m_2^2-m_1^2)\pi f_c}$	...	$\dfrac{m_2R}{2(m_2^2-m_1^2)\pi f_c}$	$\dfrac{R}{2m_1\pi f_c}$	$\dfrac{m_2(1-m_2^2)R}{2(m_2^2-m_1^2)\pi(f_2-f_1)}=\dfrac{1}{4\pi^2f_m^2C_{43}}$	...

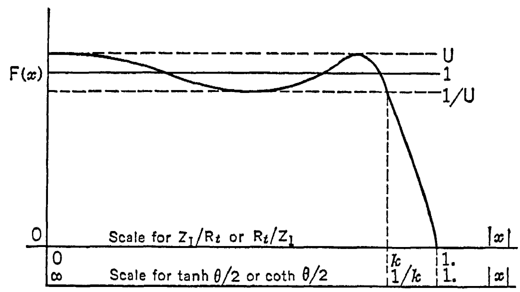
C_{12}	....	$\dfrac{m_2(1-m_2^2)}{2(m_2^2-m_1^2)\pi f_c R}$	$\dfrac{1}{2m_1\pi f_c R}$	$\dfrac{m_2}{2(m_2^2-m_1^2)\pi f_c R}$	....	$\dfrac{m_2(1-m_2^2)}{2(m_2^2-m_1^2)\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{12}}$
C_{22}	$\dfrac{m_1(m_1+m_2)}{2m_2\pi f_c R}$	$\dfrac{m_2}{2\pi f_c R}$	$\dfrac{m_1(m_1+m_2)}{2m_2(1-m_1^2)\pi f_c R}$	....	$\dfrac{m_1(m_1+m_2)}{2m_2\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{22}}$	$\dfrac{m_2}{2\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{22}}$
C_{34}	....	$\dfrac{m_2(1-m_1^2)}{2m_1(m_1+m_2)\pi f_c R}$	$\dfrac{1}{2m_2\pi f_c R}$	$\dfrac{m_2}{2m_1(m_1+m_2)\pi f_c R}$	....	$\dfrac{m_2(1-m_1^2)}{2m_1(m_1+m_2)\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{34}}$
C_{44}	$\dfrac{m_2^2-m_1^2}{2m_2\pi f_c R}$	$\dfrac{m_1}{2\pi f_c R}$	$\dfrac{(m_2^2-m_1^2)}{2m_2(1-m_2^2)\pi f_c R}$	....	$\dfrac{m_2^2-m_1^2}{2m_2\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{44}}$	$\dfrac{m_1}{2\pi(f_2-f_1)R} = \dfrac{1}{4\pi^2 f_m^2 L_{44}}$
m	$m_1 = \sqrt{1-\dfrac{1}{x_{\infty 1}^2}}$; $m_2 = \sqrt{1-\dfrac{1}{x_{\infty 2}^2}}$; $m_2 > m_1$					
Z_{I_1} (at 1-1′ terminals)	$R\sqrt{1-x^2}$	$\dfrac{R[1-x^2(1-m_1^2)]}{[1-x^2(1-m_2^2)]\sqrt{1-x^2}}$	Same as (1)	Same as (2)	Same as (1)	Same as (2)
Z_{I_2} (at 2-2′ terminals)	$\dfrac{R[1-x^2(1-m_2^2)]\sqrt{1-x^2}}{[1-x^2(1-m_1^2)]}$	$\dfrac{R}{\sqrt{1-x^2}}$	Same as (1)	Same as (2)	Same as (1)	Same as (2)
tanh θ	$\dfrac{j(m_1+m_2)\,x\sqrt{1-x^2}}{1-x^2(1+m_1m_2)}$					
x	$\dfrac{f}{f_c}$		$-\dfrac{f_c}{f}$		$\dfrac{f^2-f_m^2}{f(f_2-f_1)}$; $f_m^2 = f_1f_2$	
x_∞^2	$\dfrac{f_\infty^2}{f_c^2}$; $f_c < f_{\infty 1} < f_{\infty 2}$		$\dfrac{f_c^2}{f_\infty^2}$; $f_{\infty 2} < f_{\infty 1} < f_c$		$\left[\dfrac{f_\infty^2-f_m^2}{f_\infty(f_2-f_1)}\right]^2$; $f_{-\infty 2} < f_{-\infty 1} < f_m < f_{\infty 1} < f_{\infty 2}$; $f_{\infty 2}f_{-\infty 2} = f_m^2 = f_{\infty 1}f_{-\infty 1}$	

FIG. 16. Design Information for Terminating Sections Which Have Two Image Impedance Controlling Frequencies

the image impedance depends on the choice of the parameters, m_1 and m_2. Since the usual objective is to obtain an impedance which approximates closely to a constant resistance, these parameters may be determined most easily by the method described in the next section.

25. TCHEBYCHEFF TYPE CHARACTERISTICS

In the preceding two sections, attention has been directed at the building blocks that make up a composite filter. These elementary units can be assembled to give a great variety of characteristics. The characteristics desired in a filter design depend, of course, on the particular use for which the filter is intended. Many applications include the requirement that the image impedance should be substantially constant over a prescribed portion of the pass band. Another frequent specification is that the attenuation exceed a given value at all frequencies more than a certain distance beyond the cutoff. An equivalent statement of these requirements is: (1) the image impedance function, Z_I/R_t, should approximate to unity over a preassigned interval of the pass band; (2) the image transfer function, tanh θ, should approximate to unity over a preassigned interval of the attenuating band.



$$x = \frac{f}{f_c} \text{ for low-pass filter}$$

$$x = -\frac{f_c}{f} \text{ for high-pass filter}$$

$$x = \frac{f^2 - f_m^2}{f(f_2 - f_1)} \text{ for band-pass filter}$$

FIG. 17. The Tchebycheff Type of Approximation to a Constant Value

A desirable approximation is one for which the maximum departures of the function from unity are the minimum. This is known as the Tchebycheff approximation, proposed originally by W. Cauer for this application. In this type of approximation, as illustrated in Fig. 17, the function winds about unity in such a way that all the maximum values (U) are equal, and all the minimum values ($1/U$) are also equal. Unity is the geometric mean of the maximum and minimum values. The number of intersections with unity is determined by the degree of the function.

ATTENUATION CHARACTERISTIC. The image transfer function, tanh ($\theta/2$), of a symmetrical filter is written in terms of the frequency variable, $1/x$. The critical frequencies and the constant multiplier are so chosen that the function is constrained to lie between the limits U and $1/U$, in the frequency range from $1/k$ to infinity, as shown in Fig. 17. The intersections with unity correspond to peaks of infinite attenuation; the equal values of minimum attenuation, A_0, are equal either to 2 arg tanh U, or 2 arg coth U. For a given number of filter sections, there is a definite relation between k and A_0.

This relation is presented in chart form in Fig. 18. The parameter, N, is equal to the number of transfer constant controlling factors in the image transfer function. The number of m-derived low- and high-pass filter sections, Figs. 8 and 9, is equal to $(N + 1)$; the number of m-derived band-pass sections, Figs. 10a and 10b, is $[2(N + 1)]$. The horizontal scale of this chart is spread out considerably for values of k near unity by using the variable α defined by the equation $k = \sin \alpha$. The data required for making up this chart may be obtained from the following set of computations:

$$\alpha = \arc\sin k \tag{1}$$

$$2\epsilon = \frac{1 - \sqrt{\cos \alpha}}{1 + \sqrt{\cos \alpha}} \tag{2}$$

$$q = \epsilon + 2\epsilon^5 + 15\epsilon^9 + \cdots \tag{3}$$

$$A_0 \doteq 20(N + 1) \log \frac{1}{q} - 6.0 \quad \text{(in db)} \tag{4}$$

This relation, derived by S. Darlington, gives a result good to within 0.1 db provided A_0 is greater than about 6 db. A table of log q vs. α is contained in *Funktionentafeln* by Jahnke and Emde. This may be used in place of (2) and (3).

It is not likely that arbitrarily chosen values of A_0 and α will lie on one of the curves. Thus, a compromise must be made between minimum attenuation and the interval of the attenuating band that is covered. For example, if a filter is to be designed to attenuate frequencies above $x = 1.10$ by about 60 db, a choice must be made among the possibilities: $A_0 = 52$ db, $\alpha = 65°$, $1/k = 1.10$, $N = 2$; $A_0 = 60$ db, $\alpha = 58°$, $1/k = 1.18$, $N = 2$;

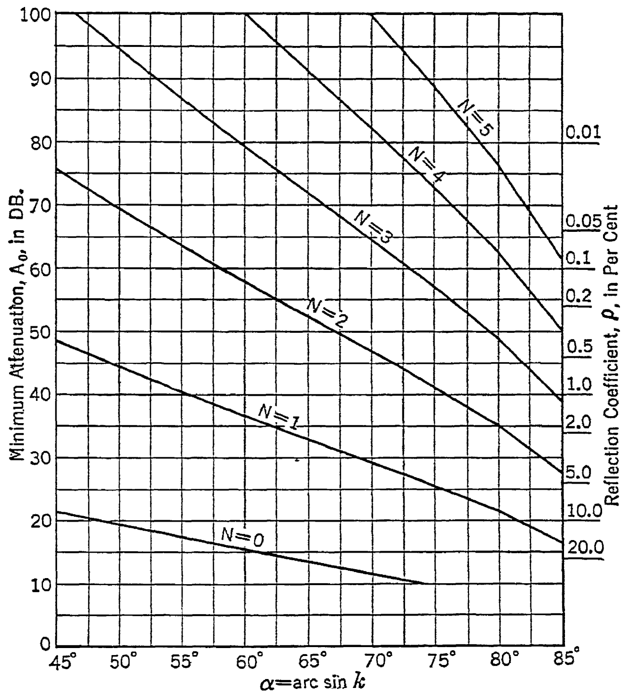


FIG. 18. Design Chart for Determining the Number, N, of Frequency Control Factors Required for a Specified Reflection Coefficient or Minimum Attenuation

$A_0 = 60$ db, $\alpha = 73°$, $1/k = 1.045$, $N = 3$. It is customary to design to a value of minimum attenuation which is 6 db greater than the desired minimum insertion loss. This allows for 3 db reflection gain at each end of the filter, which is the maximum value that may be realized.

The design formulas for the element values are expressed in terms of the impedance level, R, the cutoff frequency or frequencies, and the parameter, m. At this point it is assumed that the first two quantities are known. The values of m are determined from the frequencies of infinite attenuation. They are given by:

$$\frac{1}{x_s} = k \operatorname{sn}\left[\frac{(2s-1)K}{2(N+1)}, k\right]; \qquad s = 1, 2, \cdots, N+1$$

where sn denotes an elliptic sine function of modulus k, and K is the complete elliptic integral of the first kind. Each choice of the index s specifies an attenuation peak at the frequency corresponding to x_s.

The evaluation of the sn function can be performed by means of elliptic integral tables. However, a preferred method is to use the approximation:

$$\frac{1}{x_s} = \sqrt{k}\left[\frac{1 - u^r - u^{2-r} + u^{2+2r}}{1 + u^r + u^{2-r} + u^{2+2r}}\right]$$

where

$$2u = \frac{1-\sqrt{k}}{1+\sqrt{k}}$$

$$r = \frac{2s-1}{2(N+1)}; \qquad s = 1, 2, \cdots, N+1$$

The actual peak frequencies used in the definition of m in Figs. 8, 9, 10 are obtained from the normalized frequency for each type of filter from the following relations:

(a) low-pass $\quad f_\infty = f_c x_s$

(b) high-pass $\quad f_\infty = \dfrac{f_c}{x_s}$

(c) band-pass

$$f_\infty = f_m[Dx_s + \sqrt{1 + (Dx_s)^2}]; \quad (>f_2)$$

$$f_{-\infty} = f_m[-Dx_s + \sqrt{1 + (Dx_s)^2}]; \quad (<f_1)$$

$$D = \frac{f_2 - f_1}{2f_m}$$

$$f_m = \sqrt{f_1 f_2}$$

The band-pass attenuation characteristic is geometrically symmetrical about the mid-band frequency. That is, the attenuation at a frequency, f_x, above the upper cutoff is the same as the attenuation at the frequency $f_{-x} = f_m^2/f_x$ below the lower cutoff. Instead of designing the four-element band-pass sections, Fig. 10, it is more convenient and economical to use the six-element sections described in Fig. 14. Two half-sections must be joined together to realize the performance predicted by the above design method. It is to be noted that the parameter, m, is expressed directly in terms of x_s.

IMAGE IMPEDANCE CHARACTERISTIC. The image impedance function is written in the form Z_I/R_t in terms of the frequency variable x. The terminating resistance is denoted by R_t. The critical frequencies and the constant multiplier R are chosen in such a way that the function remains within the limits $U = R$ and $1/U = 1/R$ over the frequency range from zero to k, as shown in Fig. 17. Perfect match points correspond to the intersections with unity. Because of the reciprocal nature of the departures from unity, it is evident that the maximum departures of the magnitude of the reflection coefficient from zero are all equal over the approximation interval.

The relation between the reflection coefficient, ρ, and the pass-band interval, k, within which the reflection coefficient does not exceed the prescribed limit is presented in Fig. 18. (The scale chosen for the reflection coefficient is not a convenient one. However, it can be easily transformed into the db scale in accordance with the relation $A_0 = 20 \log (100/\rho)$, where ρ is in per cent, and A_0 in decibels.) The parameter N is equal to the number of impedance controlling factors in the expression for the image impedance. For the constant-K sections, Fig. 13, $N = 0$; for the usual m-derived sections, Fig. 14, $N = 1$; and for the "two-control" terminating sections shown in Fig. 16, $N = 2$.

The element values for the terminating sections depend on the cutoff frequencies, the impedance level, and the parameter m. It is assumed here that the cutoff frequencies are known. The impedance level is specified by:

$$\frac{R}{R_t} = \frac{1+\rho}{1-\rho} \quad \text{or} \quad \frac{1-\rho}{1+\rho}$$

where R_t is the terminating, or load, resistance, and ρ is the absolute value of the reflection coefficient (not in per cent). The first relation is used for a mid-series type impedance when N is even, and for a mid-shunt type when N is odd; the second relation is used for the other two possibilities.

The values for m are determined in terms of the frequencies of infinite attenuation. These are specified by:

$$\frac{1}{x_t} = \operatorname{sn}\left(\frac{tK}{N+1}, k\right); \qquad t = 1, 2, \cdots, N$$

The following approximation is usually more satisfactory for evaluating these peak frequencies than the use of a table:

$$\frac{1}{x_t} = \frac{1}{\sqrt{k}}\left[\frac{1 - u^v - u^{2-v} + u^{2+2v}}{1 + u^v + u^{2-v} + u^{2+2v}}\right]$$

where

$$2u = \frac{1-\sqrt{k}}{1+\sqrt{k}}$$

and

$$v = \frac{t}{N+1}; \qquad t = 1, 2, \cdots, N$$

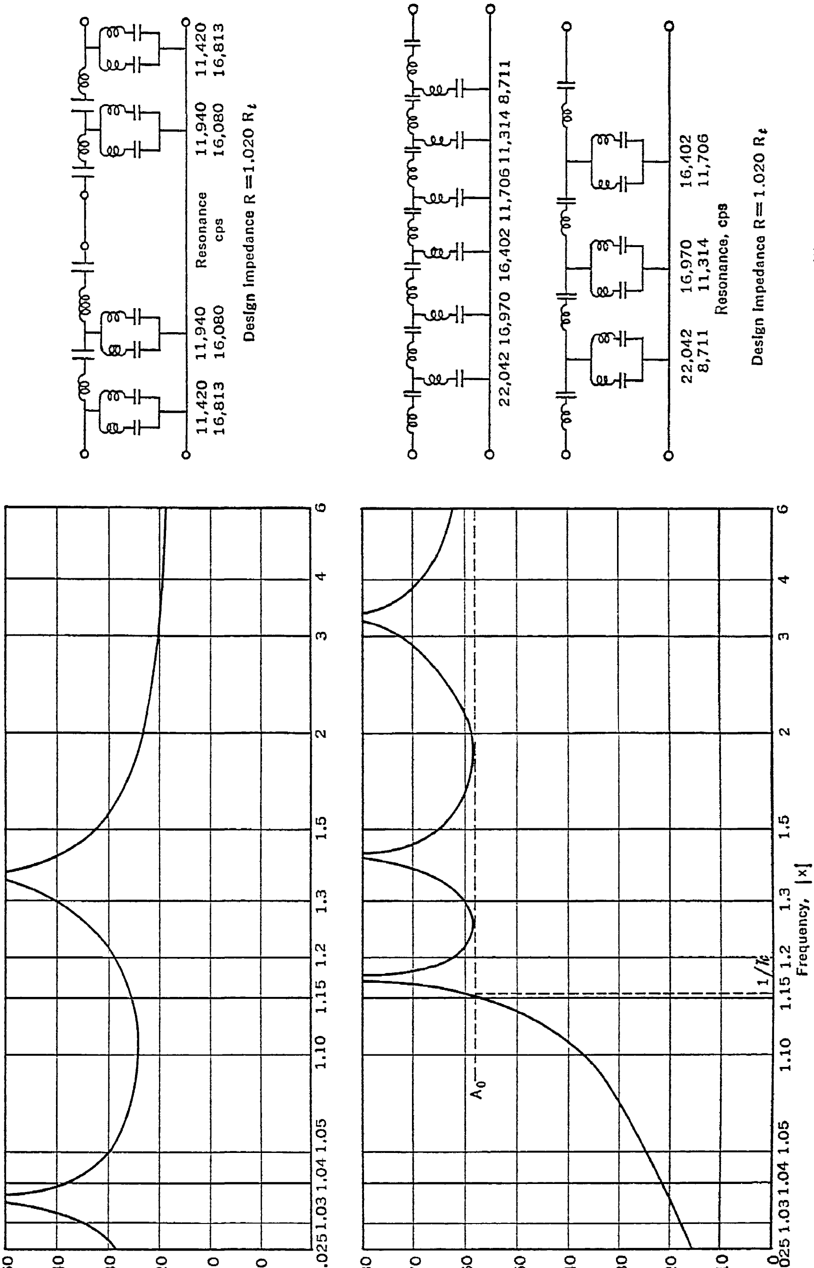
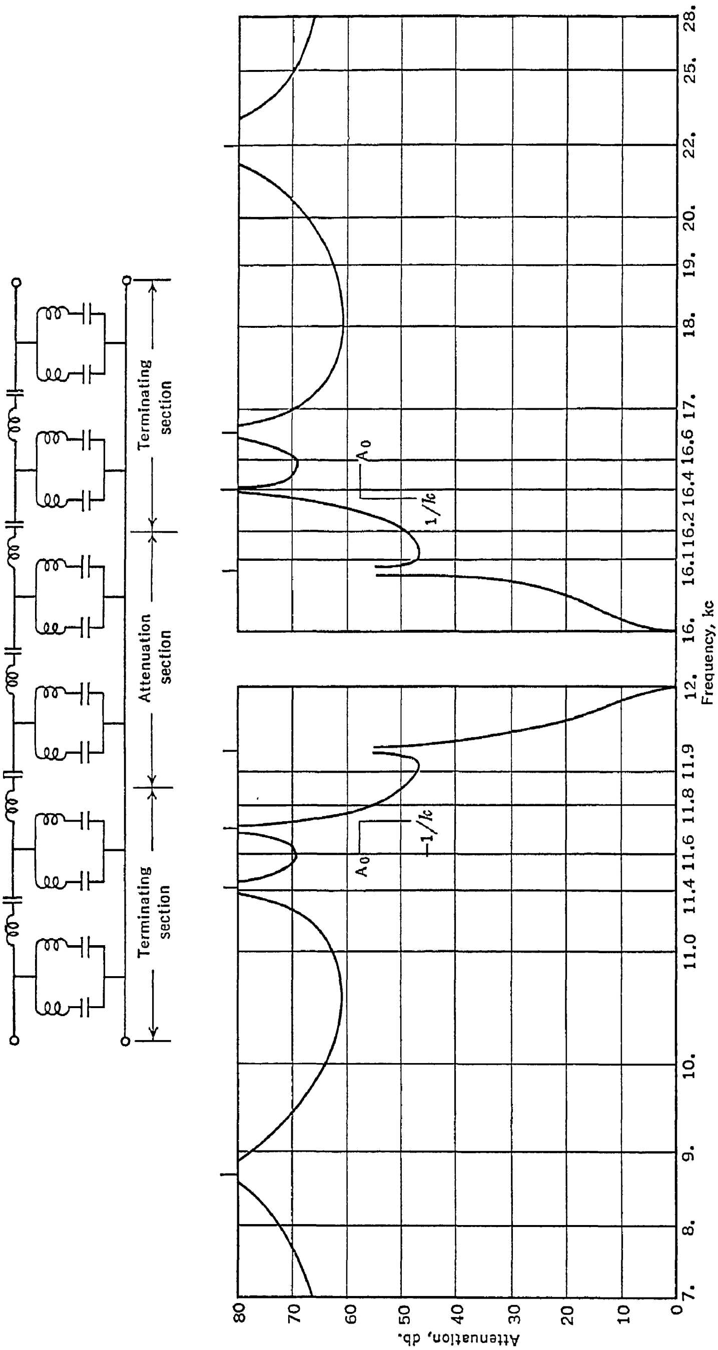


FIG. 19. Illustrative Design of a Band-pass Filter



(c)

FIG. 19 (*Continued*). Illustrative Design of a Band-pass Filter

Illustration. The application of this technique to the design of a band-pass filter serves to clarify some of the details of the procedure. It is assumed that the lower cutoff is 12 kc and the upper cutoff is 16 kc. The minimum insertion loss should be at least 52 db above 16.4 kc, and below 11.6 kc. The reflection coefficient should be less than 1 per cent between 12.1 and 15.9 kc.

The terminating sections are designed first, so that the attenuation which they contribute may be deducted from the 58 db requirement (allowing 6 db for reflection gain). From the curve, $N = 2$, on Fig. 18, it is seen that a 1 per cent reflection coefficient can be realized over the portion of the pass band corresponding to $\alpha = 76°$ by using the sections shown in Fig. 16. The actual frequency range lies between the limits $f_+ = 15{,}931$ and $f_- = 12{,}050$ corresponding to $x = k = \sin 76° = 0.9703$, as computed from:

$$f = f_m[\pm Dx + \sqrt{1 + (Dx)^2}]$$

The plus sign gives the frequency above mid-band, and the minus sign the frequency below mid-band. The quantity D is the relative bandwidth equal to $(f_2 - f_1)/2f_m$. The distance between the upper and lower frequencies is proportional to the absolute band width, i.e., $(f_+ - f_-) = (f_2 - f_1)x$.

The impedance level of the section, for $\rho = 0.01$, is found to be $R = 1.020R_t$. For precise results, a value for the reflection coefficient, superior to that read from Fig. 18, may be calculated from the equation $\rho = 2q^{N+1}$. The image impedance characteristic is similar to the curve displayed in Fig. 17 with $U = 1.020$ and $k = 0.9703$. The negative half extending from zero to minus one, corresponding to the interval from mid-band to the lower cutoff, is a duplicate of the part shown.

The final step in the design is the evaluation of the peak frequencies by means of the approximation given above, with the index $v = 1/3$ and $2/3$. The results are $x_1 = 1.0351$ and $x_2 = 1.3484$; the corresponding values of the parameter are $m_1 = 0.25838$ and $m_2 = 0.67082$. The element values are obtained from Fig. 16. An appropriate configuration is shown in Fig. 19*a*, as well as the attenuation characteristic of the two terminating sections, which is obtained by doubling the result calculated from the formula in Fig. 16.

That part of the complete filter which is between the terminating sections supplies the balance of the required attenuation, i.e., the difference between 58 db and that shown in Fig. 19*a*. This unit may be designed by fitting together the attenuation characteristics of four *m*-derived sections (Fig. 10); several trials may be necessary. It is interesting to see how the design process under discussion leads directly to a satisfactory choice for these intermediate sections.

The chart, Fig. 18, is entered with $A_0 = 58$ db. This attenuation can be obtained with three sections at frequencies above and below those corresponding to $\alpha = 60°$. The limits are found to be $f_+ = 16{,}357$ and $f_- = 11{,}738$. The values for the peak frequencies are computed from the approximation given above, with the index $r = 1/6, 3/6, 5/6$. They are $x_1 = 1.1742$, and $x_2 = 1.4142$, and $x_3 = 3.333$; the corresponding values of the parameter are $m_1 = 0.52415$, and $m_2 = 0.70711$, and $m_3 = 0.95393$. The attenuation characteristics of the three sections are calculated from the formula given in Fig. 14 (this formula applies to a half-section so that the result must be doubled). Figure 19*b* is a sketch of the complete characteristic. Two of the possible configurations are shown; the first consists of sections chosen from Fig. 10, and the second is taken from Fig. 14 (an equivalent shunt branch is used).

The sum of the characteristics, Figs. 19*a* and *b*, is far in excess of the attenuation required. It is observed that the terminating section has a peak at $x_2 = 1.35$. This is close to the peak, $x_2 = 1.41$, of one of the attenuation sections. Consequently, this latter section may be omitted from the intermediate part of the complete filter without affecting greatly the minimum attenuation of 58 db. There is a considerable increase in attenuation between the cutoffs and the limit frequencies, $\pm 1/k$, owing to the first peak in the terminating sections. However, a large part of this surplus will be lost because of dissipation in the components. The overall attenuation is plotted on an arc sin frequency scale in Fig. 19*c*.

26. MAYER'S THEOREM

The effect of uniform dissipation in the elements on the characteristics of a reactive network can be estimated readily by means of Mayer's theorem. This states that:

$$A_d \doteq A + \frac{\omega}{Q} \cdot \frac{dB}{d\omega}$$

$$B_d \doteq B - \frac{\omega}{Q} \cdot \frac{dA}{d\omega}$$

A and B represent the real and imaginary parts, respectively, of a network characteristic in the absence of dissipation. A_d and B_d are the corresponding quantities when dissipation is present. As usual, ω is 2π times the frequency. The average dissipation is:

$$\frac{1}{Q} = \frac{1}{2}\left(\frac{1}{Q_L} + \frac{1}{Q_C}\right)$$

where Q_L is the average $\omega L/r$ for the coils, and Q_C is the average $\omega C/g$ for the condensers. Frequently, the condensers are considered relatively non-dissipative, so that $Q = 2Q_L$.

Since these relations come from a Taylor's series development of the network function, they are usable only in the regions where the function is well behaved. For example, if A and B represent image attenuation and image phase shift, the approximation fails in the neighborhood of the cutoffs and the attenuation peaks; on the other hand, if A and B represent the insertion loss and insertion phase shift, the approximation holds everywhere except near the loss peaks; a similar remark obtains for driving point impedance, current ratio, voltage ratio, etc., functions.

These formulas lead to some useful general conclusions. They indicate that, to a first approximation, the change produced in the real component of a network characteristic by dissipation is proportional to the slope of its imaginary component, and vice versa. It is particularly interesting to note the effect in respect to characteristics which approach the ideal. For example, the approximation to an ideal filter is designed to have a constant (zero) attenuation and a phase shift which varies linearly with frequency over most of the pass band. It is to be expected that the phase characteristic will be changed very little by dissipation since the slope of the attenuation is zero. On the other hand, the attenuation characteristic will be affected when dissipation is present. However, over the frequency range for which Q is proportional to frequency, the attenuation will be constant and proportional to the phase slope. Consequently, the filter introduces a flat loss but does not introduce distortion.

REFERENCES

Bartlett, A. C., *The Theory of Electrical Artificial Lines and Filters.* John Wiley (1930).
Bode, H. W., A General Theory of Electric Wave Filters, *J. Math. and Phys.*, Vol. 13, 275–362 (November 1934).
Bode, H. W., and R. L. Dietzold, Ideal Wave Filters, *B.S.T.J.*, Vol. 14, 215–252 (April 1935).
Cauer, W., *Siebschaltungen.* V.D.I. Verlag G.m.b.H., Berlin (1931).
Cauer, W., New Theory and Design of Wave Filters, *Physics*, Vol. 2, 242 (April 1932).
Darlington, S., Synthesis of Reactance 4-Poles, *J. Math. and Phys.*, Vol. 18, 257–353 (September 1939).
Foster, R. M., A Reactance Theorem, *B.S.T.J.*, Vol. 3, 259–267 (April 1924).
Guillemin, E. A., *Communication Networks*, Vol. II. John Wiley (1935).
Zobel, O. J., Theory and Design of Uniform and Composite Electric Wave Filters, *B.S.T.J.*, Vol. 2, 1–46 (January 1923).

RADIO ANTENNAS

By J. C. Schelleng *

GENERAL FUNCTION AND DESCRIPTION. The antenna is the means of coupling between the medium of propagation and the transmitter or receiver. Its purpose is to convert power into outgoing electromagnetic waves, or to extract power from incoming waves. These two processes are reciprocal, but other factors such as discrimination against static introduce new requirements which differentiate a good transmitting from a good receiving antenna.

Not only should an antenna be an efficient converter between radiant and guided energy, but it should also be most effective in the direction of the station with which it is communicating. At the receiving end, directivity performs two useful functions. The first is that of impressing on the early stages of the receiver a relatively intense signal so as to override the inherent circuit or tube noises. The second and often the more important is that of discriminating against electrical disturbances in the medium, such as atmospheric noise, thereby increasing the ratio between signal and noise components. If the atmospherics should arrive from the direction of the signal, then, of course, no advantage would result from this means. When, on the average, they arrive with equal intensity from all directions, the advantage is considerable. If it happens, as in communication between Europe and America, that a very small part of the total noise arrives from the direction of the desired signal, a very great advantage is realized. Indeed, because of their ability to discriminate against atmospherics arriving from the rear, it is sometimes the practice to use as receiving antennas devices that would ordinarily be relatively poor radiators.

Practical antennas differ widely. Perhaps the simplest is a straight vertical wire as shown in Fig. 1*a*, *b*, *c*, and *d*. For the lowest frequencies used in radio the length is a very small fraction of a wavelength, but with higher frequencies the vertical wire may exceed a half-wavelength. The dotted lines indicate roughly the distribution of current. Several antennas may be arranged to form a directive array as shown in Fig. 1*l*. An antenna one-half wave long (Fig. 1*g*) is known as a half-wave linear antenna.

When the frequency is low (long waves), the antenna may take the form of an L or a T, a hundred or so feet high and several hundred long (see Fig. 1*e* and *f*). The horizontal portion is known as the *flat-top*, the vertical as the *lead-in*. Frequently a large coil of wire is used as an aerial as shown in Fig. 1*k*. It is known as a loop.

---

* In this revision a considerable amount of material from the article Radio Antennas (7–57) appearing in the 1936 Edition of the Pender-McIlwain *Handbook*, and written by G. C. Southworth, has been used.

In the highest decade of the radio spectrum (in 1949 the practical limit seems to be of the order of 30,000 megacycles per second, wavelength 1 cm), the low-frequency technique associated with radiation "from the outside" of the conductors has become relatively

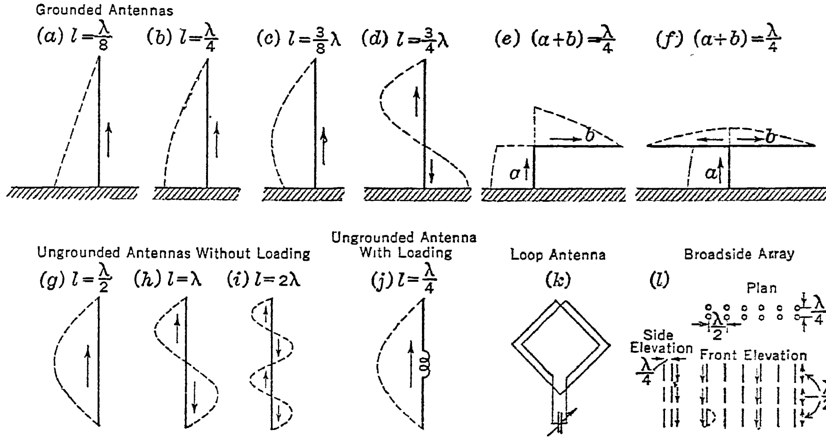


FIG. 1. Representative Forms of Antennas

difficult owing to the small size of the wavelength, and antenna designers tend to draw their inspiration from wave-guide principles and from optics. Typical antennas now are found to employ wave-guide apertures or horns, parabolic reflectors or lenses, as well as many

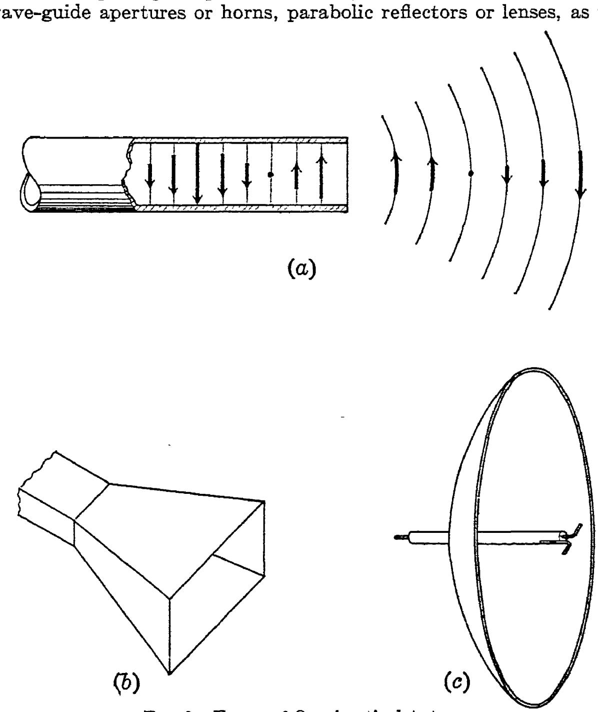


FIG. 2. Forms of Quasi-optical Antenna

of the older techniques. Figure 2*a* illustrates radiation from the open end of a circular wave guide, 2*b* shows a pyramidal horn formed by tapering from a small rectangular wave guide, and 2*c* shows a linear half-wave antenna at the focus of a paraboloidal antenna.

CLASSIFICATION. Table 1 is an attempt to classify the many diverse forms of antennas which have found use. Ignoring distinctions between transmitting and receiving, it gives one type of classification based on general form rather than specific application. In spite of a degree of artificiality it covers most of the common types. The relatively non-directional appear early in the list; the highly directional are found at the end.

Table 1. Classification of Antennas

Linear Conductor Antennas
- L, T, umbrella, multiple-tuned electric dipoles, "vertical radiators," whip, trailing-wire, vertical wire, horizontal wire, wave antenna, loops, etc.

Wave-guide Antennas
- Small wave-guide apertures:
  - Open-end wave guide.
  - Slot antennas.
  - Various reflector or lens feeds, etc.
- Large wave-guide apertures: Horns.
- Dielectric antennas: Polyrods.

Quasi-optical Devices: Antennas Using Reflectors or Lenses for Collimating
- Spherical optics: Point sources ("dipoles" or wave-guide apertures) in conjunction with spherical reflectors or lenses.
- Cylindrical optics: Line sources (arrays, reflectors, lenses) in conjunction with cylindrical reflectors or lenses.

Arrays or Combinations of Above Devices
- Long-wire antennas.
  - Wave antenna.
  - V-antenna.
  - Rhombic antenna, etc.
- Linear arrays.
  - Broadside array.
  - End-fire array.
  - Fishbone.
  - Musa.
  - Turnstile.
  - Clover leaf, etc.
- Curtain arrays.
  - Franklin.
  - Pine-tree.
  - Chireix-Mesny.
  - Sterba.
- Arrays of: Horns; quasi-optical devices; wave-guide apertures.

RADIATION, ABSORPTION, AND RECIPROCITY. In general, the radiation from an antenna can be calculated if the tangential magnetic and electric fields subsisting over any closed surface containing the antenna are known. The tangential magnetic field measures linear density of an equivalent electric current in the surface at right angles to the tangential field; likewise the tangential electric field measures a linear density of "magnetic current." General formulas (beyond the scope of this article) exist for calculating from these the radiated field (S. A. Schelkunoff, Reference 1, 9.1-7, 9.1-9 and 9.1-10). They are necessary, for example, in calculating the radiation from an electromagnetic horn. In many other cases, including practically all antennas of older types where the energy prior to radiation was guided near the outside of conductors rather than through hollow guides, this procedure simplifies: the closed surface now may be taken as the surface of the conductor of the antenna itself and the electric currents of the equivalent sheet become the usual antenna currents, while the "magnetic currents" become zero owing to the disappearance of a tangential electric field at the surface of the conductor. The fact that in the lower-frequency ranges the electric current in conductors can be conveniently measured by ammeters gives a special importance in those ranges to formulas in which the antenna current is assumed to be known.

An analogous process occurs in reception, in which equations might be set up for integrating the total effect on the receiver due to the electric and magnetic distributions over a closed surface containing the antenna, but, since these fields are in part reradiation associated with the response of the receiving antenna, the problem is more involved than that of the transmitting antenna and it is common to invoke the law of reciprocity instead. As applied to radio wave propagation through a simple linear medium (excluding nonlinear circuit elements and the ionosphere), this law says that, if there is a single zero-impedance generator in the transmitter and a zero-impedance ammeter in the receiver, the generator and the ammeter may be interchanged without affecting the current measured. An alternative expression of the law is: if a constant-current generator in the transmitter produces a reading in an infinite-impedance voltmeter in the receiver, the generator and the voltmeter may be interchanged without affecting the reading. These statements say nothing explicitly about the power transfer, but the following one does: if the internal resistance R of a generator having zero reactance is matched to the transmitting antenna, and if the receiving antenna is matched to a load whose impedance is R, then the power transfer will not be affected by interchanging the generator and the load. It follows from any of these statements that *a given antenna has the same directional pattern for receiving as for transmitting.*

27. PRINCIPLES OF LINEAR CONDUCTOR ANTENNAS

What takes place in even a simple transmitting antenna is a matter of such great complexity that a rigorous description is beyond the scope of this article; nevertheless it will be well to recognize the nature of the problem. One aspect of it is the calculation of external fields, assuming a knowledge of the distribution of current in the conductors.

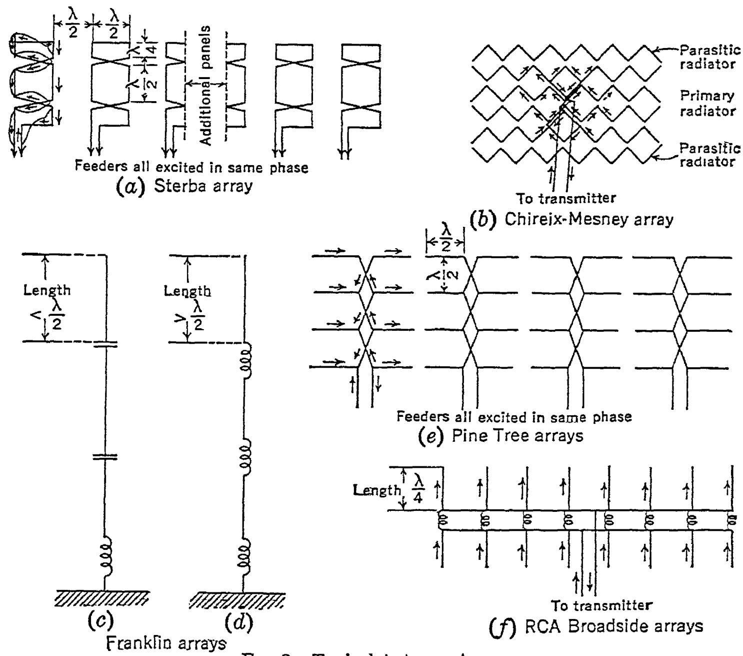


FIG. 3. Typical Antenna Arrays

This problem is completely solved by integrating for each point in space the retarded effects of currents at different parts of the antenna. To the extent that we are satisfied with the assumed current distribution, a satisfactorily complete formal solution is always possible. Many results of great engineering usefulness have been so obtained; the usual assumption is that the current is distributed sinusoidally, that the standing-wave pattern is formed by equal waves traveling oppositely with the velocity of light (3×10^8 meters per second), with current nodes at open ends of conductors. This means that the standing wave will have minima at intervals of one-half wave, with maxima at intermediate points, the instantaneous current being oppositely directed either side of a minimum, as shown in Fig. 1*d*, 1*h*, and 1*i*. In certain problems it is convenient to regard the radiation as issuing from certain centers, such as the midpoint or maximum between adjacent minima. On this view a small radiator may have a center of radiation much as an extended body has a center of mass. For sections of conductor whose length is small compared with a wavelength the current distribution may be substantially uniform, as in the vertical lead of a large flat-top antenna, or in a loop small compared with a wavelength. There is a considerable field where the engineer may employ this concept without apology to the mathematician.

There are other cases where he needs to watch his step carefully, as for example in the antenna shown in Fig. 3*a* where for qualitative purposes the standing wave is shown along the folds of the wire. The existence of current minima in this antenna was checked experimentally by E. J. Sterba, but he found that, as the antenna was extended by adding more sections at the top, the lower minima became less definite and ceased to be nulls. This is an extreme case perhaps beyond mathematical solution, but it illustrates the matter.

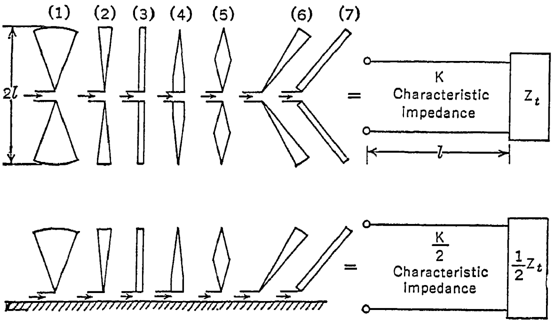


FIG. 4. The input impedance of a conical antenna of any size is equal to the input impedance of a uniform transmission line with a certain "output" impedance Z_t. The input impedance of a thin antenna of any shape is similarly represented, except that the characteristic impedance is variable.

A second aspect of this theoretical problem is the actual determination of the currents and voltages that exist on an antenna having specified the applied or external forces. S. A. Schelkunoff has given a general theory of certain forms of antenna which is especially attractive to the engineer (see Reference 1). In ordinary transmission lines, the distribution of a linear charge density is the same as that of voltage, but this is not generally true in antennas. The explanation is that antennas, unlike ordinary transmission lines, support more than one mode of propagation. As far as

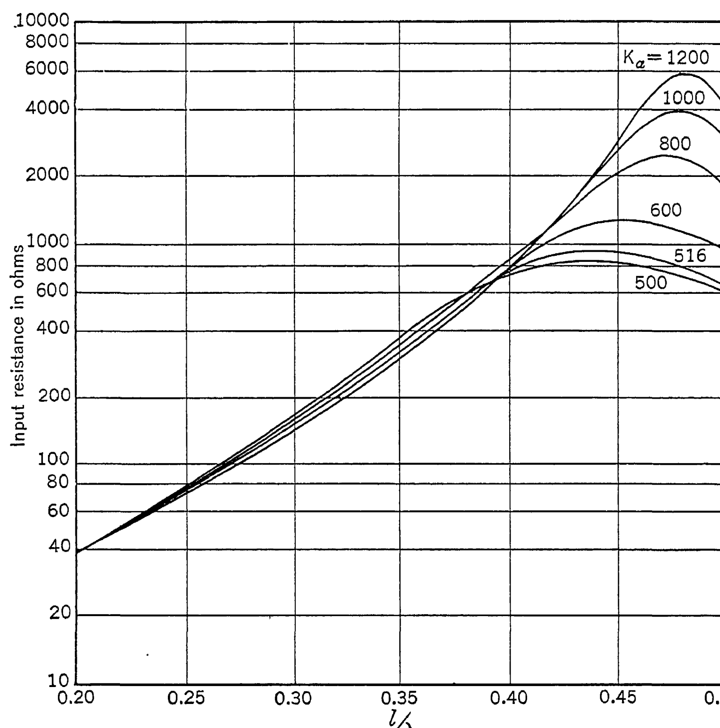


FIG. 5. The Input Resistance of Hollow Cylindrical Antennas in Free Space. For vertical antennas over a perfectly conducting ground divide the ordinates and K_a by 2.

the current associated with the principal mode is concerned (the mode that we are thinking of when we draw the oversimplified current-distribution curves which we have just been discussing), radiation is strictly an end effect: "It is permissible to think that a wave emerging from a generator in the center of an antenna is guided by the antenna until it reaches its boundary sphere passing through the ends of the antenna and separating the antenna region from external space; at the boundary sphere some energy passes into external space and some is reflected back—a situation existing at the juncture between two transmission lines with different characteristic impedances." This will be clearer by reference to Fig. 4, taken from Schelkunoff, where the "antenna region" or sphere is replaced by a transmission line having length equal that of the antenna, the line having appropriate characteristic impedance K and terminating impedance Z_t. The legend makes the figure self-explanatory.

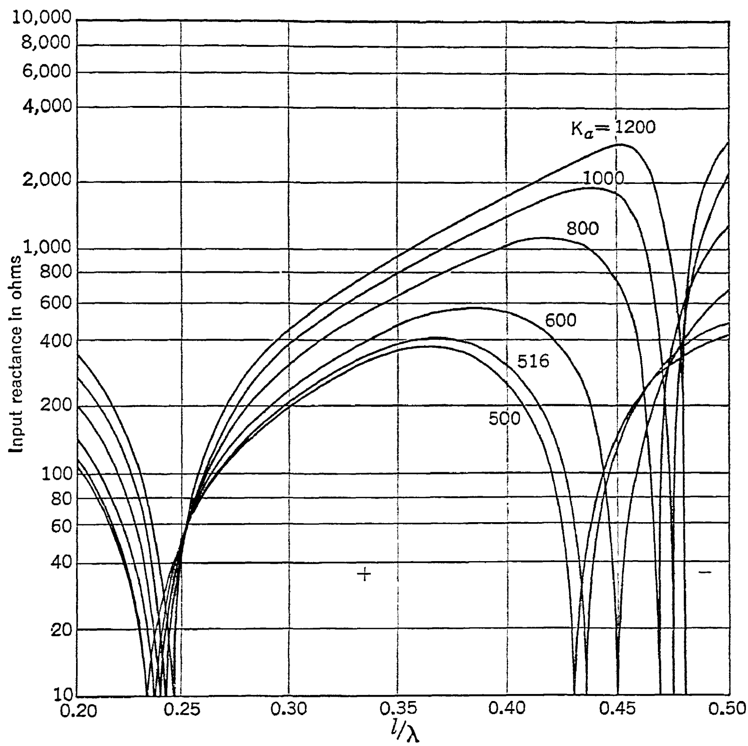


FIG. 6. The Input Reactance of Hollow Cylindrical Antennas in Free Space. For vertical antennas over a perfectly conducting ground divide the ordinates and K_a by 2.

The input resistance and reactance of perfectly conducting cylindrical antennas in free space as given by Schelkunoff are plotted in Figs. 5 and 6, the parameter K_a, the characteristic impedance, being given in Fig. 7. These curves were obtained on the assumption that the current flowing over the edge at the top of the antenna is zero. For $K_a < 700$ there is a small current over the edge which, if included in the calculations, would increase the maxima of the input resistance and reactance by a small percentage. Note the points where input reactance is zero, and the deviation of the lengths from 0.25λ and 0.5λ, as found experimentally by C. R. Englund.

In some low-frequency antennas where radiation resistance is low, the input reactance has often been calculated by regarding the antenna as a transmission line without resistance, using the formula

$$x = -\sqrt{\frac{L_1}{C_1}} \cot (\omega l\sqrt{L_1C_1}) \tag{1}$$

where L_1 and C_1 are inductance and capacitance per unit length, l the total length, and ω the angular frequency. Such a line is resonant when $l = \lambda/4$, $3\lambda/4$, etc., and antiresonant when $l = \lambda/2$, λ, etc. If a lumped inductance is connected in series with the long antenna assumed, resonance will occur when

$$\omega L - \sqrt{\frac{L_1}{C_1}} \cot (\omega l \sqrt{L_1 C_1}) = 0$$

The effect of capacitive loading may be found in a similar manner.

When an antenna of this kind does not exceed a quarter-wave in length it may be roughly considered part of a resonant circuit made up of lumped inductance and capacitance as

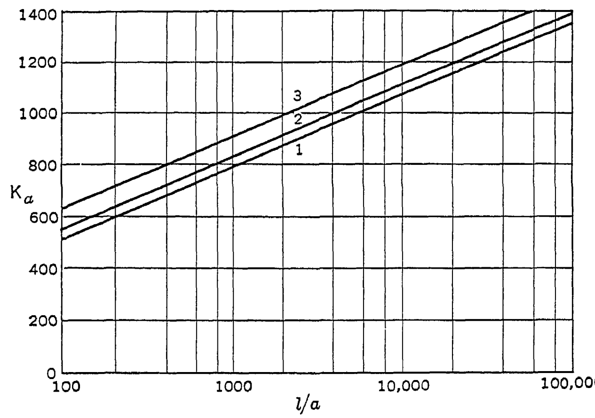


FIG. 7. The Average Characteristic Impedance: (1) Cylindrical Antenna, (2) Spheroidal Antenna, (3) Antenna of Rhombic Cross-section. l is the length from middle of antenna to the ends; a is the maximum radius of conductor.

follows: $L_e = lL_1/3$ henrys and $C_e = lC_1$ farads together with whatever loading may have been added. The shorter the antenna the better is the approximation.

RESISTANCE OF ACTUAL ANTENNAS. Antenna resistance is the quotient of the mean power supplied to the antenna divided by the mean square of the current referred to a specified point of the antenna. It thus includes a component associated with the useful radiation of power and others related to undesirable losses in the conductors, ground, etc. Radiation resistance is thus the quotient of the mean radiated power divided by the mean square of the current referred to the specified point, and radiation efficiency is the ratio of radiation resistance to total resistance.

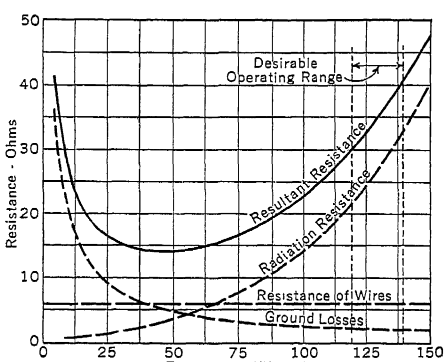


FIG. 8. Components Which Go to Make Up the Total Resistance of a Simple Antenna

Figure 8 shows the general way in which radiation resistance, with the losses in the ground and in the conductors, varies with frequency. An efficient antenna being one in which radiation resistance predominates, the desirable operating range is well to the right in the figure. At the lowest of radio frequencies, however, economic factors makes high efficiencies unrealizable, and efficiencies of a few per cent, or less, are representative. In the example shown in Fig. 8, the wire resistance is substantially independent of frequency, but in general this is not true.

The radiation resistance of a straight vertical wire of infinitesimal diameter referred to the point where it connects through a coupling or load impedance to a perfect ground varies from zero to 36.6 ohms as its length increases from zero to a quarter-wave (half the value given in Fig. 5). A convenient approximation for the radiation resistance provided that the actual height is well below $\lambda/4$ is $160\pi^2h^2/\lambda^2$, where h is an effective vertical length of the radiator and λ is the wavelength in the same units. If, for example, we consider a vertical wire $\lambda/4$ long (which we have just stated is too long for this approximation), the assumed sinusoidal current distribution makes $h = 2/\pi \cdot \lambda/4 = \lambda/2\pi$, and radiation resistance of 40 ohms is indicated. Comparison with the correct value of 36.6 ohms indicates that for antennas shorter than $\lambda/8$ the approximation will usually be acceptable.

When the operating frequency is low it is uneconomic to construct a vertical wire even approaching a quarter-wave in length. Under these circumstances it is customary to combine a rather large flat-top with a moderate vertical lead, in order to hold costs to the minimum. This leads, however, to low radiation resistance and so requires that other resistances be kept correspondingly low. Thus, if the radiation resistance should be as low as 0.05 ohm, an elaborate counterpoise or other ground system would be necessary to keep the losses within reasonable bounds.

ANTENNA IMAGES. A portion of the wave radiated from an antenna is "reflected" by the ear that points some distance away. To an observer located at a considerable distance the total radiation appears to be made up of two components, one which arrives directly from the antenna itself and the other which appears to be coming from a virtual antenna located below the surface of the earth. The latter, sometimes known as an image antenna, behaves in different ways depending on the soil adjacent to the antenna. Figure 9 pictures the relation between the effective currents in real and image antenna, assuming the earth to be perfectly reflecting.

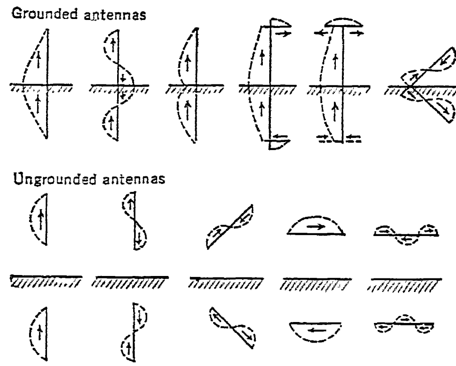


FIG. 9. Images of Representative Forms of Antennas

FIELDS ASSOCIATED WITH AN ANTENNA. The wave radiated from an antenna appears as two fields, (1) an electric field $\mathcal{E}$ which may be measured in volts per meter, and (2) a magnetic field which may be specified in amperes per meter. These components are so inseparable as to be regarded as two aspects of the same thing. They are perpendicular to each other and to the direction of wave propagation and at any point of observation are in the same phase, reaching zeros and maxima at the same time. For a fundamental discussion on this subject the reader is referred to Section 5, articles 26–28.

The electric field produced at a distance of a few wavelengths from an antenna may be calculated by means of the general formula applying to an elemental doublet

$$\mathcal{E} = \frac{60\pi}{dc} f \delta l I \cos \omega \left(t - \frac{d}{c} \right) \cos \theta \tag{2}$$

where $\mathcal{E}$ = the field intensity of the wave measured in volts per meter; $I \cos(\omega t + 90°)$ = current flowing in the wire in amperes; f = frequency of the current in cycles per second; λ = wavelength corresponding to frequency f; $\omega = 2\pi f$; t = time in seconds; d = distance to the antenna in meters; c = velocity of light = 2.998×10^8 meters per second; δl = the elementary length of wire or doublet from which radiation takes place—it is measured in the same units as λ; and θ = angle of elevation of point at which the field is desired measured relative to a plane perpendicular to the conductor δl. In the formulas in this section, any other unit of length can be used, provided it is used consistently in $\mathcal{E}$, λ, d, c, A, etc. The meter, however, is preferred.

Equations for the fields radiated by antennas of various shapes, such as given just below, are obtained by integrating the above doublet expression over the entire conductor system, having due regard for the distribution of the current. The fields at great distances from antennas are usually much less than those calculated by these formulas. (See Section 10.)

The *voltage induced* in an incremental length of conducting wire by a passing wave may be found by either of two methods. The one starting with the electric field of the wave is used the more. That component of the length lying parallel to the electric vector in the wave front is multiplied by $\mathcal{E}$ measured in volts per unit length to give the required emf. This same voltage may also be derived from the simple dynamo concept of the number of lines of magnetic force cut per second; it will be realized, however, that this is not another component of induced voltage but merely another approach to the same one made possible by the identities expressed by Maxwell's equations.

The instantaneous radiated power flowing through each unit of area perpendicular to a plane wave front may be calculated by the expression

$$P = 0.00265\mathcal{E}^2 \quad \text{watt per square meter} \tag{3}$$

(See also Sect. 3 art 28 eq. [16].)

If eq. (3) is used for calculating the power picked up by a single wire, it must be assumed that power is absorbed from a section of the wave front extending to about a quarter of a wavelength either side of the conductor.

The field intensity at a horizontal distance d from a *vertical grounded wire*, length h, carrying a uniform current I is approximately

$$\mathcal{E} = \frac{120\pi fhI}{dc} \quad \text{volts per meter} \tag{4}$$

This applies to the vertical lead to an antenna having a relatively large flat top, the antenna being located over a perfectly conducting earth.

The corresponding value for a coil or *loop* located *in free space* and having dimensions small compared with the wavelength is

$$\mathcal{E} = \frac{120\pi^2 f^2}{dc^2} ANI \quad \text{volts per meter} \tag{5}$$

where N is the number of turns and A is the area in square meters.

The effective vertical length of a *loop in free space* is given by

$$h = \frac{2\pi AN}{\lambda} \quad \text{meters} \tag{6}$$

The field intensity for a simple *vertical quarter-wave* antenna grounded at its lower end is approximately

$$\mathcal{E} = \frac{60I}{d} \quad \text{volts per meter} \tag{7}$$

The effective vertical length of a *grounded-quarter-wave* antenna is

$$h = \frac{\lambda}{2\pi} \quad \text{meters} \tag{8}$$

In all these cases the current is to be measured at the point in the antenna where this quantity is a maximum.

For formulas for field strength in terms of radiated power, see Section 10, Articles 20–24 of this handbook.

Figure 10 gives the directional diagram of a simple electric or magnetic radiating element (a short wire or a small loop). It is worth noting that the same diagram applies to both.

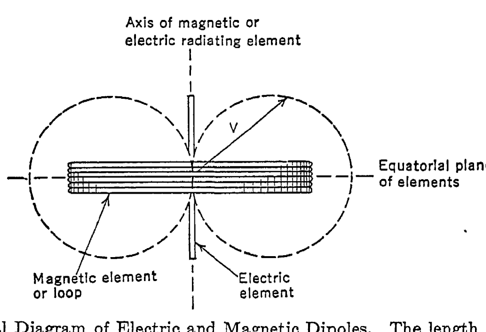


Fig. 10. Directional Diagram of Electric and Magnetic Dipoles. The length of the vector V is proportional to signal intensity.

The directional diagrams for both the quarter-wave and half-wave antennas as well as for an antenna 0.62 wavelength long are given in Fig. 11. It is to be noted that for this latter a pronounced spurious lobe is formed. This continues to grow as the antenna is further lengthened, thereby leading to a considerable amount of high-angle radiation and possibly also to fading. The field laid down in the horizontal direction is for a given power calculated to be a maximum when the length of the antenna is about 0.62 wave. (See reference 2.) The half-wave antenna together with its earth image is roughly equivalent to two collinear equiphased radiators and therefore is a special case of arrays discussed below. Antennas a half-wave or so long have become widely used in broadcasting. See "Antennas for Medium-frequency Broadcasting," article 31.

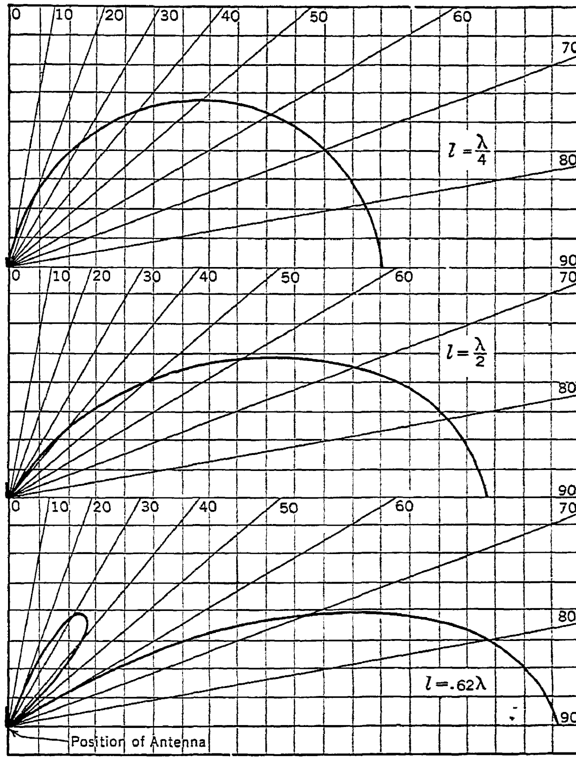


Fig. 11. Vertical Plane Directional Diagram of a Single Vertical Antenna of Various Lengths. (Only one-half of the total vertical section is shown.)

28. PRINCIPLES OF DIRECTIVITY

An antenna which radiates or receives with uniform efficiency within a range of direction is said to be *non-directional* within those limits; when it favors a given direction, on the other hand, it is *directional*. Thus, a vertical wire is non-directional in azimuth, though it is directional in elevation. Ignoring non-coherent radiation, such as light, it may be said that, for pure sinusoidal currents, a completely non-directional antenna does not exist. It is, however, a useful concept to define the directivity of actual antennas in terms of this imaginary non-directional source. Thus, the *absolute gain* is commonly defined as

$$G = \frac{P}{P_0} \tag{9}$$

where P is the power flow per unit area in the plane linearly polarized wave which the antenna causes in a distant region (usually in the direction of maximum radiation), and P_0 is the power flow per unit area which would have been produced if all the power had been radiated equally in all directions. Where gain is referred to any other standard, it will be specifically mentioned.

One special form of radiator, the current element short in comparison with a half-wave, has frequently been used for reference in discussing directivity. It is, in fact, as non-directional as a simple radiator can be. Actually, however, it has a power gain of 1.5 over the standard described in the preceding paragraph.

Though considerable directivity can at least in theory be accomplished with an antenna whose largest dimension is small compared with one wavelength, it is not far from the truth to say that all practical high-gain antennas depend primarily on having current distributed over dimensions of several wavelengths. This statement applies both to linear-conductor antennas and to quasi-optical devices.

The requirements of directivity are various. It may be necessary to limit antenna effectiveness to a certain azimuth, to a certain elevation, or to both. A narrow beam may be required in azimuth coupled with a wider one in elevation, as in some transatlantic antennas used in the range of high frequencies; or a relatively sharp beam may be wanted in elevation with little or no directivity in azimuth, as in television broadcasting; or sharpness may be needed in azimuth with a specified variation of intensity in elevation, as in the "cosecant" antennas of radar, etc.

The special distribution of current needed for directivity may be provided by arrays of similar smaller radiating elements or by combinations of dissimilar elements. These building blocks may be half-wave wires, loops, flat-top antennas, and so forth. The distribution may be continuous: for example, an electromagnetic horn may be thought of as an array of an infinite number of infinitesimal radiators.

An almost unlimited number of combinations is possible, since the directive effects produced depend not only on the relative positions and spacings of the various units but also on the amplitudes and phases of their currents as well. (See reference 3.)

Perhaps the best-known arrangement is that of a number of identical parallel antennas arrayed laterally along a straight line. Usually the element antennas carry equal in-phase currents. This produces the strongest signal in a direction perpendicular to the line of the array and is thus known as a *broadside array.* An example is shown in Fig. 1*l*, above. In some cases the phasing of the currents is progressively delayed from antenna to antenna to correspond exactly with the delay in that direction due to finite wave velocity. Such an antenna is called an *end-fire* array because the radiation is most intense along the line of the array.

EFFECTIVE AREA OF ANTENNAS. A wave incident upon a receiving antenna may be thought of as a stream of energy possessing a certain power per unit of cross-sectional area. If the receiver load is coupled to the antenna so as to abstract the maximum power available, then the ratio of this maximum power to the power incident on the antenna per unit area is defined as the effective area of the receiving antenna. A somewhat oversimplified view is that the antenna presents this area to the energy stream and canalizes the corresponding power flow into the receiver. An excellent treatment of relations to be discussed in this section will be found in an article by H. T. Friis and W. D. Lewis. (See reference 4.)

The effective area of a receiving antenna being by definition proportional to its power gain, we may make the general statement that the ratio power-gain divided by effective area has the same constant value for all antennas. Numerically it turns out that this ratio is:

$$\frac{\text{Gain}}{\text{Eff. gain}} = \frac{G}{A_{\text{eff.}}} = \frac{4\pi}{\lambda^2} \tag{10}$$

an important relation in antenna theory. It can be shown that it has a useful interpretation in transmitting as well as in receiving: this same area then measures that broadside "uniformly excited" area which would give the same transmitting gain as does the actual transmitting antenna, the excitation being unidirectional (e.g., as when a reflector is used) and the dimensions of the area being large compared with a wavelength. Hence, a lossless transmitting antenna in which the radiation is associated with a large uniformly excited area would, as a receiving antenna, make available to the receiver all the energy intercepted by its actual area.

A very useful and simple *free space transmission law* (see reference 4) results by application of these concepts to wave propagation between antennas of effective area A_T and A_R (the subscripts refer to transmitter and receiver). The power delivered to the receiver is then P_R, which equals

$$\frac{\text{Total power}}{\text{Area of sphere of radius } d} \cdot G_T \cdot A_{\text{eff. (receiver)}} = \frac{P_T}{4\pi d^2} \cdot \frac{4\pi A_T}{\lambda^2} \cdot A_R$$

giving

$$P_R = P_T \frac{A_T A_R}{\lambda^2 d^2} \tag{11}$$

The fact that the numerical constant turns out to be unity recommends this formula to the memory.

29. DIRECTIVITY OF LINEAR CONDUCTOR ANTENNAS

ARRAYS. Of the many directive patterns that may result from the various spacings and phasings of two antennas, two are of especial interest. In the first the two sources are separated in space by one-fourth of a wavelength, and in phase by one-fourth of a period. This arrangement, which is sometimes known as a unidirectional couplet, gives a cardioid pattern as shown in Fig. 12*a*, where the unit antennas are vertical half-wave elements. As compared with a single element it effects a power gain of about 2 (3 db). In the other arrangement the two elements are spaced one-half wavelength and are driven in phase (see Fig. 12*b*). This also gives a theoretical gain of about 2 (actually somewhat greater). The two arrangements may be combined as in Fig. 12*c* to give a total gain of 4 (6 db). It is convenient to regard the two antennas at the rear as reflectors for those ahead. Directional effects such as these are used practically not only to increase signal in some desired direction but also to minimize its interfering effects in others.

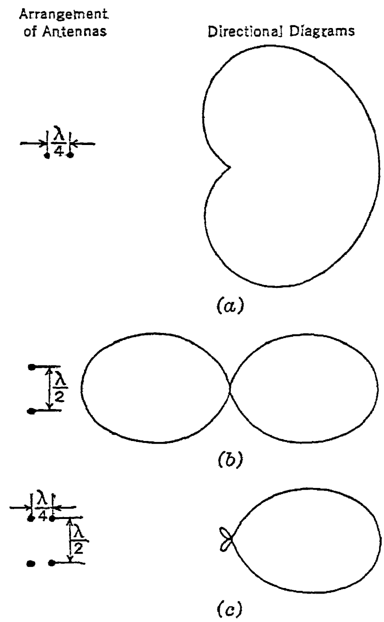


FIG. 12. Horizontal Directional Diagrams. (*a*) Unidirectional couplet. (*b*) Two equiphased antennas spaced one-half wave-length. (*c*) Two equiphased couplets.

Increased directivity may be obtained by adding couplets to the arrangement shown in Fig. 12*c*. The resulting increase is indicated by Fig. 13. Although it is often most convenient in practice to utilize spacings of one-half wave in the array front, any spacing up to about 3/4 wave may be used. For spacings less than 0.6λ it is the total length or aperture of the broadside which is the important criterion of gain, the gain variation due to spacing being inappreciable. Figure 14 facilitates determining the gain ratio of such arrays. The *aperture* there referred to is about one-half wave greater than the number of wavelengths measured between extreme conductors of the array. The reason for this rule is that the equivalent area of a thin wire properly coupled to the terminal is finite (see paragraph following eq. [3]).

When an array is formed by stacking similar units in tiers in the vertical direction, added directivity is provided at some angle from the vertical, commonly at 90°, that is, in the horizontal plane. The units may, for example, be the vertical elements previously described, and if the elements of a broadside array are so arranged the antenna resembles a curtain. For vertical elements the improvement obtained by adding a small number of tiers or stacks is less than that achieved by the lateral arrangement. (See reference 5.)

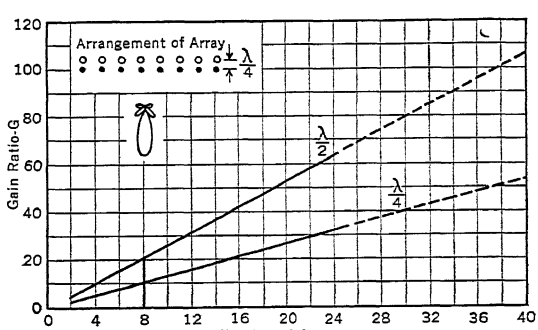


FIG. 13. Variation of Directivity with Number of Couplets Placed in Horizontal Array

Simple but approximate rules for unidirectional broadside arrays are as follows:

1. The gain ratio of a large array of vertical couplets extending both laterally and vertically (and including a reflecting curtain) follows the general rule for gain $G = 4\pi A/\lambda^2$ or, in terms of a short current element as standard, $G_d = 8/3 \cdot \pi A/\lambda^2 = 8.4A/\lambda^2$, where the effective length is taken one-half wave greater than the length between extreme conductors (see paragraph following

eq. [3]), and the height as $\lambda/2$ times the number of tiers or stacks. This assumes no appreciable gap between tiers. For a single-tier antenna $G_d = 10A/\lambda^2$.

2. Doubling the length or the height, or adding the "reflector" curtain to the front curtain, adds 3 db to the gain. Note the exception already indicated, however, that in going from one to two tiers the increase is only about 2 db.

Arrays of this type have achieved considerable importance both in long-distance transmission using short waves (high frequencies) and in medium-frequency broadcasting. In the former use very considerable power gains have been employed, varying up to more than 100. In the latter the gain has been valuable in extending coverage, but the most important aim has been the suppression of signals in certain directions at night in order to avoid interference with other stations. Thus the directivity of broadcast transmitting antennas has two aims essentially corresponding to those mentioned in the first section with respect to directivity in reception, viz., increase of signal and reduction of interference.

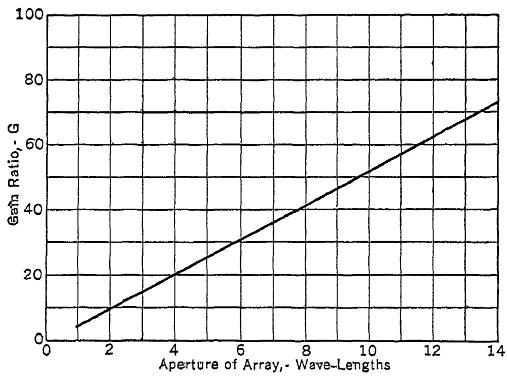


FIG. 14. Graph for Predicting Directivity of Arrays of Simple Half-wave Antennas. (Aperture is expressed in wavelengths and is one-half wavelength greater than the horizontal distance between the extreme outside antennas.)

The main problem in medium-frequency broadcasting which leads to the use of directional antennas is the difficulty of giving local coverage without causing interference at distances of a few hundred to a few thousand miles with other transmitters using the same frequency. This long-distance interference is propagated by reflection from the ionosphere, and therefore not only azimuth but also elevation must be considered. Regardless of the height of the reflecting layer, the azimuth of waves between two points lies along the great-circle path with considerable consistency, and an antenna which directs a minimum of signal at all elevations in the vulnerable azimuth is usually the most desirable solution. Sharp nulls in the elevational directive pattern are thus to be avoided if possible in view of the variability in the heights of the reflecting layers. Often, however, a host of practical considerations requires a compromise solution. (See reference 6.) Commonly, also, the problem is complicated by the existence of more than one vulnerable direction. The necessity of having the correct relative phases in the unit antennas has required the development of techniques for controlling, measuring, and maintaining phases in practical installations. (See reference 7.)

DIRECTIONAL CHARACTERISTICS OF LONG WIRES. When the length of a wire carrying a high-frequency current is progressively increased, it breaks up into oscillating sections as was shown in Fig. 1*i*. This gives rise to radiation along certain preferred directions in which in-phase components prevail. Figure 15 shows the directional patterns for certain representative cases. It will be noted that the lobe designated as No. 1 ear becomes progressively sharper and approaches the axis of the wire. At the same time smaller lobes designated as No. 2, 3, and 4 ears are formed.

Several long wires each having characteristics of this kind may be so combined as to give the arrangement as a whole very useful directional properties. In general, this is accomplished by choosing arrangements that enhance the main lobe and at the same time discourage the spurious lobes. The so-called tilted wire, folded wire, and rhombic antennas are based on this principle. (See reference 8.)

EFFECTS OF SOIL AND TERRAIN ON DIRECTIVITY. The directive effects described above assume that the array is divorced from any influence of the earth. In practice, of course, this is not true. If the earth were perfectly conducting the array could be so elevated as to make the image effect add to that of the array, thereby giving added gain and a maximum intensity along the surface of the earth. These ideal conditions are seldom attained in practice. At the frequencies at which directive antennas are most used, there is a substantial refractive effect in addition to absorption that together tend to distort the vertical directive characteristic.

Figure 16 shows the calculated directional distortion imposed by imperfectness of earth conductivity on both a horizontal and a vertical receiving doublet for a representative case. It is to be noted that the effect is less marked with a horizontal half-wave antenna than with a vertical half-wave, and in both forms it is such as to cut down very materially the intensities of waves along the horizontal. It fortunately happens that distant signals

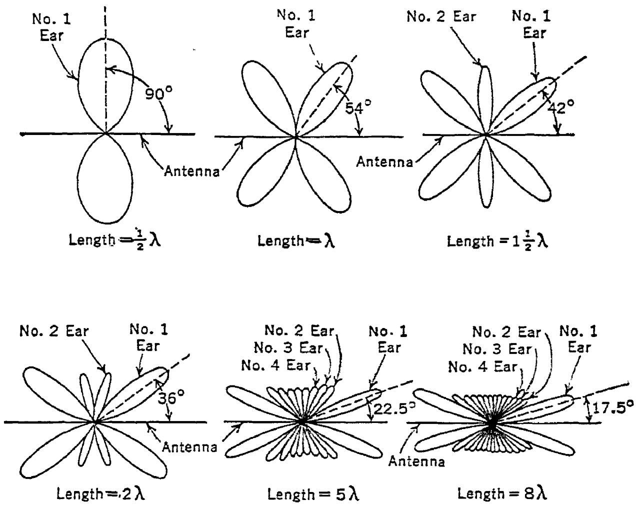


FIG. 15. Directional Diagrams of Isolated Wires of Various Lengths

arrive at an appreciable angle above the horizon, so that such devices are still very effective. While this angle for transmission and reception of short waves may, for short distances, be nearly 90°, for long distances it varies, say, from 30° for the lower frequencies to very small angles for the higher frequencies. It would appear, therefore, that the distortion of the vertical directive pattern caused by a soil of finite resistivity might constitute a definite limitation in working with very high-frequency stations. It is seen from Fig. 16 that horizontal antennas are inherently high-angle devices and that vertical antennas may also be high-angle devices except when located over a low-loss earth.

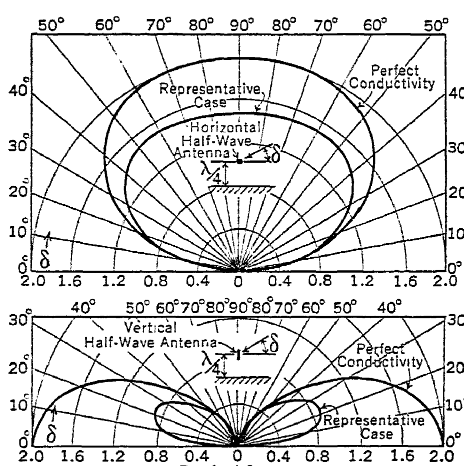


FIG. 16. Vertical Directional Diagrams of Horizontal and Vertical Half-wave Antennas as Influenced by Finite Conductivity of Earth

Advantages ranging up to 10 db (see reference 9) have been obtained by locating short-wave antennas and arrays at the tops of sharp declivities or long slopes. These gains are comparable with those of the arrays themselves and are such as may warrant considerable time in the selection of the site of a short-wave radio station. They may be explained either as due to the antenna being in a position where the field distribution is more favorable or by saying that the declivity has effectively lowered the angle of elevation of the antenna itself.

As the factors that effect vertical directivity vary markedly from point to point over the country it is difficult to present any considerable number of representative data in the space here available. However, both terrain and soil conditions are important in antenna design and should be considered when any large expenditures are to be made.

30. DIRECTIVITY OF QUASI-OPTICAL ANTENNAS AND HORNS

(See reference 4)

THE HUYGENS SOURCE. The optical concept of wave propagation according to which a wave front is considered an array of secondary sources is of great importance in analyzing the behavior of quasi-optical antennas. A formulation consistent with fundamental electromagnetics has been given by S. A. Schelkunoff (reference 1, 9.1 and 9.24). Commonly its usefulness arises in situations like that, for example, at the mouth of an electromagnetic horn, where we have reason to believe that the currents represented by the wave front predominate in ultimate effect over currents elsewhere, such as those on the outside of the horn. If we know the distribution of intensities over this aperture, Huygens' principle can be applied. In general, over this surface, polarization is distributed in both the x and y directions, but for simplicity we write the equation for only the first of these. Assuming the dimensions of the aperture great enough to give sharp directivity, we are mainly concerned with directions not far removed from the center of the main beam, which is near the axis of the reflector. Under these conditions the field at a distance r in front of the mirror is parallel to E_0 in the aperture and is given by the expression:

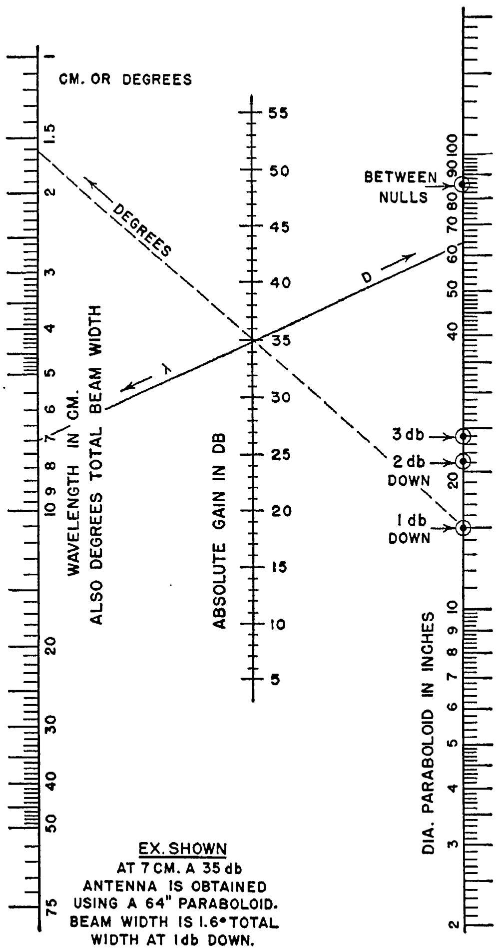


FIG. 17. Nomogram—Paraboloid Antenna Data

$$\mathcal{E}_x = \frac{i}{\lambda}\int\!\!\int \frac{\mathcal{E}_0 e^{-2\pi i r/\lambda}}{r}\, dS$$

dS being the element of area of the aperture, and λ the wavelength.

THE APERTURE. The theoretical performance of a non-dissipative antenna, which as a transmitter has a uniformly equal distribution of E_0 over its aperture, is useful for comparison with actual transmitting or receiving antennas. It can be shown from the above equation (see reference 10) that for reception its effective area equals the actual area of its aperture. In other words, it can be expected to capture from the wave all the energy "intercepted." Note, however, that antennas in reception usually cannot do this, since as a rule they do not, in transmitting, produce a constant E_0 over the whole aperture, and since, moreover, they may not be large enough compared with a wavelength to validate our assumptions. As applied to transmitting, the term effective area may be interpreted as the area of uniform excitation which would give the same field at the same distance and with the same total power as would the actual antenna.

In practice the effective area of large apertures is usually considerably less than the actual area, and the ratio is an "efficiency" factor which usually lies within the range 0.4 to 0.7, the deficiency being due primarily to the non-uniformity of intensity across the aperture. It should be observed that requirements other than gain may make it necessary to avoid uniformity across the aperture, such as the desirability of suppressing minor lobes of the directional pattern.

Other points in connection with the aperture are: (1) whether or not the amplitude is uniform, it is usually important that both phase and polarization be the same at all points in a plane perpendicular to the desired direction of transmission; (2) the width of the

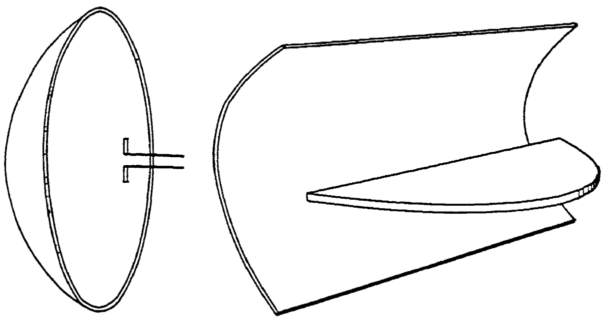


FIG. 18. Examples of Cylindrical and Spherical Optics

aperture a of a broadside antenna is inversely as the angular beam width required in the plane containing that dimension. At the half-power points the beam width is 51 λ/a in degrees for uniform illumination through a rectangular aperture and for non-uniform illumination of circular or elliptical apertures it is typically 65 λ/a. The gain and beam width of circular apertures having tapers of illumination commonly used at present are given in the nomogram, Fig. 17. When the beam width required is different in the two planes, the aperture widths are affected inversely, an elliptical shape being common. (See reference 11.)

Although the operation of most microwave directional antennas is best understood in terms of generation of a plane wave front, there are notable exceptions, such as linear end-fire arrays and polyrods.

POINT SOURCES AND LINE SOURCES. In transmitters and receivers, the power is conveyed to and from the antenna in transmission lines which are smaller than a

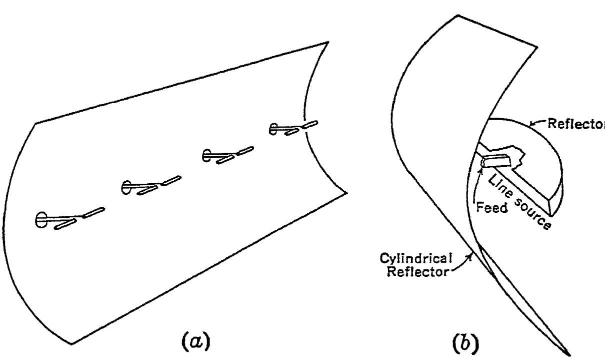


FIG. 19. Cylindrical Collimation

wavelength in cross-section, whereas in the antenna the dimensions may be of many wavelengths. The antenna must, therefore, include a distribution system. This may consist of a branching system, such as that used commonly in arrays. In most microwave antennas it consists of a "primary feed" or radiator which launches a wave, this wave then being allowed to spread in azimuth and elevation simultaneously (spherical optics), or in azimuth and in elevation successively (cylindrical optics). The latter two processes are indicated in Fig. 18.

In spherical expansion the wave must first be launched by a primary feed antenna which is basically a point source. Examples will be given later. When, on the other hand, successive cylindrical expansion is used, the wave from a point source is usually first confined between closely spaced plates and allowed to expand in a plane, and then the line of the wave front is converted from a circular arc to a segment of straight line; thereafter, the three-dimensional wave front expands from this line as a cylinder whose elements are parallel to it. These operations may be followed in the example given in Fig. 18 or 19*b*. The line source can be formed in other ways, such as a linear array of half-wave antennas, thus avoiding the first step of cylindrical expansion (Fig. 19*a*).

COLLIMATING DEVICES: REFLECTORS AND LENSES. Starting with the energy diverging from a point source, some device is needed to convert the wave front to a plane, that is, to make the emergent "rays" parallel. For this purpose reflectors or lenses are used.

(a) Parabolic Reflectors. The choice between a paraboloid and parabolic cylinders (see Fig. 18), depends on many mechanical and electrical considerations. In the past, the paraboloid has been the more used. Among the advantages sometimes claimed for it are greater electrical simplicity, lower weight and better efficiency, better directional pattern in the desired polarization, and adaptability to conical lobing or spiral scanning.

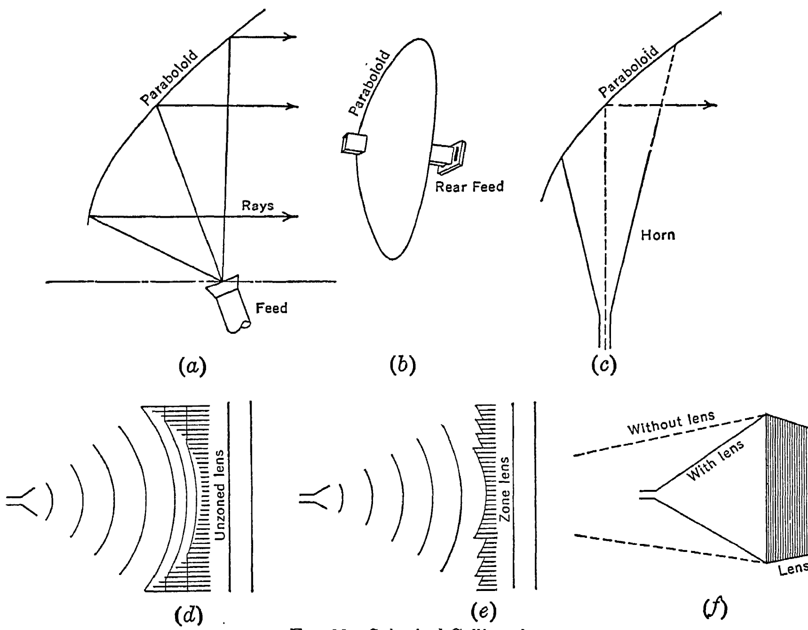


Fig. 20. Spherical Collimation

The cylinders, on the other hand, may be simpler mechanically and possess separate control of directivity in azimuth and elevation, a point of controlling importance in some applications.

For both types of parabolic reflectors, the feed is located at or near the focus, and the section used may be symmetrical about the axis or off to one side (usually the former).

(b) Lenses. Dielectric lenses can be used, and they have found considerable application in the first step of cylindrical expansion described above, i.e., in the formation of line sources, low-loss polymers being commonly used as dielectric. For spherical optics, however, such a lens becomes too massive. W. E. Kock (see reference 12) has developed other types of lenses particularly suitable for microwave use, for example that which employs metal plates instead of dielectric material. All lenses depend on having a material in which the phase velocity is different from that of surrounding space. If a plane wave were incident on a bottomless metallic honeycomb in a direction parallel to the cells, it would pass through more or less unimpeded if the frequency were above the critical frequency of the individual cells considered as wave guides. However, the phase velocity in these wave-guide cells would be greater than that of light, and the material as a whole would therefore possess a refractive index less than unity. Lenses can, therefore, be constructed by grading the depth of the cells in a manner analogous to that of optical lens practice, with this difference, that a form which causes divergence in a lens of glass will produce convergence in a lens of this cellular material; for example, a planoconcave cellular lens can be used as a collimator. If the wave is linearly polarized the cell spaces may be made indefinitely wide in the direction of the electric vector, the lens then becoming an assembly of spaced metallic plates; if in this case, however, the electric vector is made perpendicular to the plates, the wave passes through but with substantially the velocity of light, so that the device does not act as a lens. The lens can be given a stepwise reduction in thickness by means of "zoning," the "riser" of the step being that length of wave guide necessary to include one cycle less than a corresponding distance in free space. (See Fig. 20.)

HORNS. Just as the hollow wave guide is the analog of the speaking tube, so the electromagnetic horn is the analog of the acoustical horn in function as well as in appearance. It is usually the tapered extension of a metallic wave guide, as shown in Fig. 2*a* and *b* though it can be excited in other ways.

As in parabolic reflectors, the directional properties and gain of horns are determined by the excitation across the aperture, and the considerations given at the beginning of

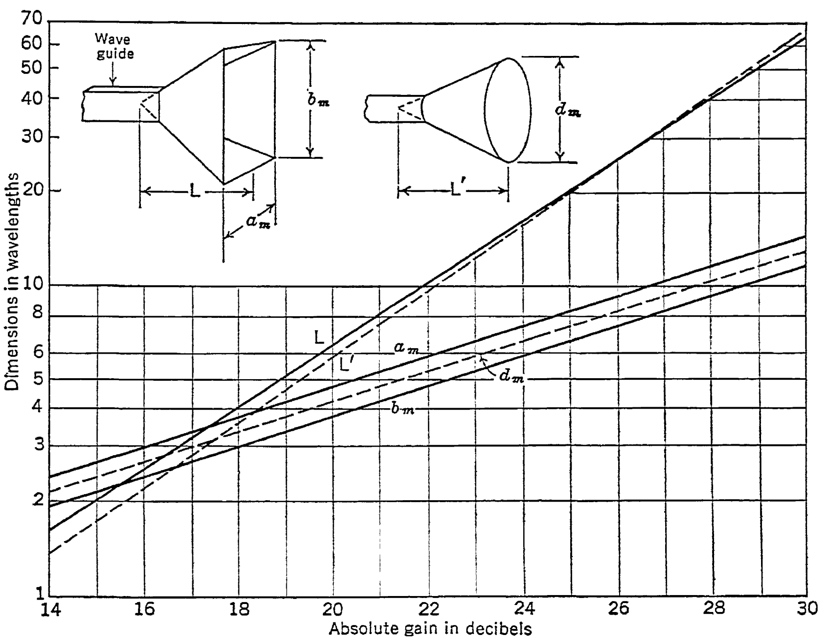


FIG. 21. Optimum Horn Data. The electric vector is parallel to the dimension b_m

this article on "The Aperture" apply approximately. The distribution of intensity in the aperture of horns tends to be uniform in the E plane and sinusoidal in the H, a condition resulting from the maintenance of the distribution natural to a small wave guide in passing from the throat through the taper to the aperture. Roughly then the half-power beam width in the E plane of large horns approximates 51 λ/a_E, and in the H plane 65 λ/a_H, provided that a, the aperture dimension, is larger than one wavelength.

The phase at the center of the aperture of a pyramidal horn tends to lead that at the edge owing to the difference in distance to the throat. For a given size of aperture this phase difference approaches zero as the length is increased, and the best length is infinity since any phase difference tends to reduce the gain. In a simple horn, therefore, the length is more likely to set a practical limit than the aperture, and for a given practical length it is often important to know the flare or aperture that will give greatest gain. This "optimum horn" can also be defined as the minimum length of horn that will give a required power gain. It does not provide greatest efficiency of aperture area but deliberately tolerates an increase in area for a decrease in length.

Figure 21 gives the essential dimensions of optimum horns, conical and pyramidal, which are taken from the article by A. P. King. (See reference 13.) The area efficiency of a large conical optimum horn is about 55 per cent; that of a very long horn of the same aperture is near 80 per cent. (See reference 13.)

Among the forms of horn are the conical, biconical, sectoral, and pyramidal. Horns can be used with lenses to avoid the undesirable phase difference across the aperture. Solid or metal plate lenses may be used. The advantage of a lens is that it very greatly reduces the length of the horn for a given aperture. (See Fig. 20.)

31. PRACTICAL ANTENNA SYSTEMS

In modern radio practice the highest frequencies are more than one million times as great as the lowest, extending as they do beyond the range from 30,000 cycles to 30,000 megacycles. We have emphasized the unity of principle within this gamut of frequency. On the other hand, in covering the field of practical antennas it is necessary to examine types in great diversity. Space prevents any attempt to be comprehensive. In this article we can consider only representative antennas which illustrate different engineering problems to be met in practice.

ANTENNAS FOR LOW FREQUENCIES. In Fig. 1*e* and *f* are shown prototypes of antennas important during the first two decades of "wireless," antennas which still are of practical importance below 1000 kc. Their common characteristic is the use of a flat-top

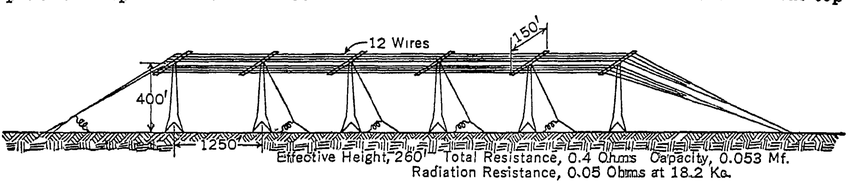


Fig. 22. Typical Multiple-tuned Antenna

and a down-lead, but these have taken on many special forms—L, T, umbrella—with single or multiple down-leads. Except in favored situations (e.g., on board a ship), it has been necessary to build more or less elaborate ground systems comprising buried wires or an overhead "counterpoise" in order to raise the radiation efficiency to an acceptable value, and even then this efficiency might at the lowest frequencies be only a few per cent. Two general methods of improving efficiency are, first, to increase the height of the towers used (increase radiation resistance), and second, to use ground systems as already stated, frequently accompanied by multiple tuning (decrease the ground losses). Figure 22 shows an Alexanderson multiple-tuned antenna having six multipled down-leads and six tuning coils; its effectiveness depends on the fact that, although the ground resistances associated with the various down-leads act as though in parallel to give a low resultant ground resistance, radiation resistance is not correspondingly reduced. Obviously structures built on so large a scale are very expensive and have to be designed with great attention to economic factors. (See reference 14.)

Figure 23 shows an antenna installation for a 150-kc land station such as has been used in communication with ships at sea.

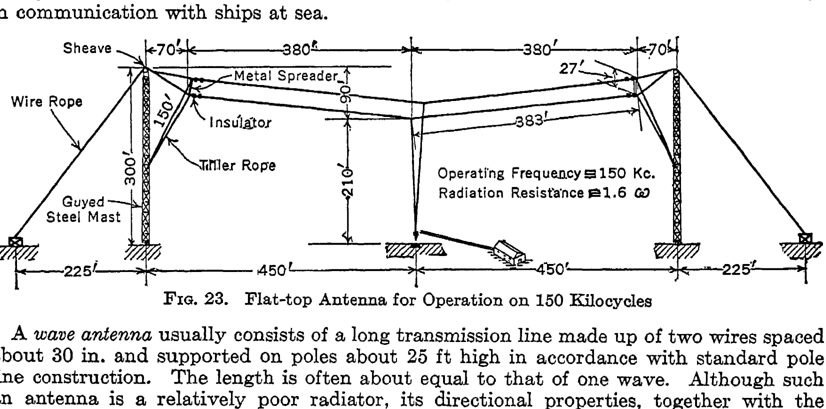


Fig. 23. Flat-top Antenna for Operation on 150 Kilocycles

A *wave antenna* usually consists of a long transmission line made up of two wires spaced about 30 in. and supported on poles about 25 ft high in accordance with standard pole line construction. The length is often about equal to that of one wave. Although such an antenna is a relatively poor radiator, its directional properties, together with the rather wide band of frequencies which it can accommodate, make it exceedingly useful in long-wave work. In particular, when used as a receiver, it is able to discriminate markedly against static arriving from other than the preferred direction of reception.

This provides a favorable ratio of signal to static. Because of its broad frequency characteristic it is possible to attach two or more receivers and simultaneously receive several frequencies. These signals must, of course, be arriving from the same general direction. The directional characteristic depends among other things on soil resistivity. Such antennas have not proved particularly effective in regions of high rainfall and high conductivity. Several wave antennas may be placed in broadside array as described above or they may be placed one back of another and sidestepped to form a staggered array.

Figure 24 shows in schematic form a wave antenna and its associated terminating network. The impedance Z is equal approximately to the characteristic impedance of the

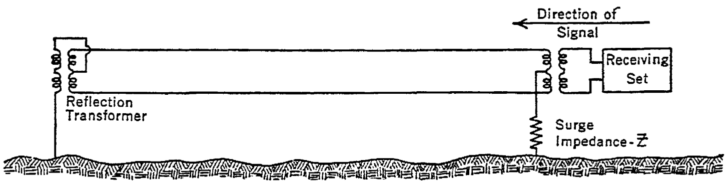


FIG. 24. Schematic of Wave Antenna. (The length may be as much as a mile, the height about 25 feet.)

antenna. This prevents reflection and renders the device essentially unidirectional. The reflection transformer shown is an ingenious means whereby the accumulated signal received between the two wires and ground may be transmitted back to a receiver located at the incident end, over the metallic circuit consisting of the two wires themselves. (See reference 15.)

ANTENNAS FOR MEDIUM-FREQUENCY BROADCASTING. Resonant antennas of the general type discussed in the previous sections (e.g., Fig. 23) have been used in broadcasting. A more common form, however, uses a tower or mast itself as the current-carrying conductor and radiator. Some of the forms which it takes, shown in Fig. 25, illustrate its basic simplicity. The self-supporting towers may be of constant cross-section or tapered to a point at the top. The masts, supported by guys sectionalized by insulators to prevent them from taking part as radiators, are commonly tapered over the lower fraction of their height to a single compression insulator and ball-and-socket joint at the base. The upper portion of the mast may be tapered toward the top, but a top without this taper is probably more common. (See reference 16.)

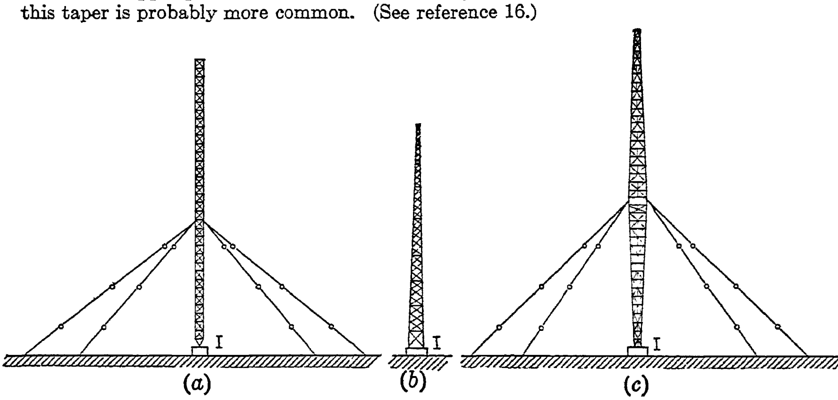


FIG. 25. Typical Forms of Broadcast Antenna

All the antennas shown in Fig. 25 are insulated at the base. There is a form called the "shunt-fed antenna" (see reference 16), however, in which the tower is connected directly to the ground network. The feed wire is connected, not at the base, but sufficiently far above it to include an appreciable tower inductance.

For daytime coverage in this frequency range the desideratum is a strong field in the horizontal direction, since waves leaving in an upward direction are ineffective either for good or for ill because they are absorbed by he ionosphere. Assuming antenna resistance to be confined to radiation and current to be distributed sinusoidally, the maximum horizontal field theoretically is obtainable when the height of the radiator is $^5/_8$ wavelengths

(225 electrical degrees, assuming that length of wave on the tower is the same as in free space). As shown in Fig. 11, however, such an antenna will have a strong minor lobe at 30° from the vertical, and at night when ionospheric absorption tends to disappear waves may thus be received strong enough to be comparable with the ground wave. Undesirable fading may then be as serious as insufficient field strength would be, particularly in the upper half of the daytime frequency range. It has, therefore, become the practice to use radiators of that height and current distribution which gives the best compromise between field strength and freedom from fading, and tower heights from 0.53 to 0.55 λ (190° to 200°) are widely used. (See reference 6.)

The considerations of directional pattern just described aim to protect the station's listeners from its own sky wave. At points beyond the normal ground-wave range are receiving sets tuned to other broadcasters operating on the same frequency, and they must be protected from interference. New minima of signal strength in the elevation plane must therefore be provided in that azimuth, and in view of the great distances of the receivers the minima must be aimed far from the vertical, typically in the order of 75°. For this purpose arrays of antennas are common; such arrays for broadcasting have been discussed above under "Directivity of Linear Conductor Antennas: Arrays." The proper design of such an array is a matter of some complication; not only is the accurate locating of towers and the specifying of their relative currents an exacting matter for calculation but also the experimental realization of the desired amplitudes and phases in the presence of the large mutual impedances which exist between towers makes it necessary to provide means for accurately adjusting measuring and maintaining the currents both in amplitude and in phase. (See reference 7.)

SPECIAL ANTENNAS FOR BROADCAST RECEPTION. Many broadcast receivers are supplied with built-in antennas which, although inefficient as compared with the "custom-built" antennas that are feasible in point-to-point work, give adequate performance, thanks to the high gains of receivers and the surplus signal laid down most of the time by powerful transmitters. In the few cases where both received field and atmospherics are below set noise advantage may be taken of large antennas. If the antenna is located where the ratio of ambient noise to ambient signal is too great, it can frequently be relocated in a quieter spot remote from the receiver and may be connected to the receiver by a shielded transmission line. This is sometimes done in apartment houses, a broad-band amplifier then being employed to permit distribution to a large number of users. In the nature of the case broadcast reception of medium frequencies does not ordinarily permit the use of much directivity. (See reference 17.)

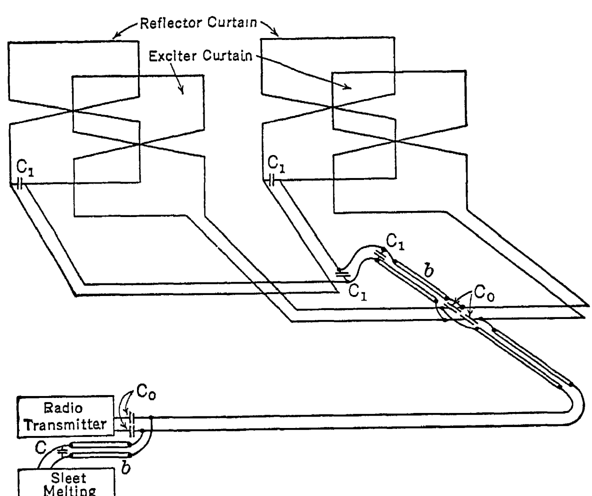


FIG. 26. Arrangement of Conductors and Impedance Matching Devices in a Sterba Array

ANTENNAS FOR HIGH FREQUENCIES (2500 TO 25,000 KC). In this band of frequencies, almost exclusively used for long distances, the wavelength is small enough to make directivity feasible. Formerly the majority of directional-antenna installations employed broadside arrays of half-wave elements operative over a relatively narrow band of frequencies. Many of these have been replaced by some form of long wire, which gives considerable directivity at moderate cost and is sensibly aperiodic so that it may be operated at several frequencies simultaneously. Resonant antennas still find use, however, for example, where space considerations do not permit the more extended aperiodic types.

Figure 3 shows schematics of various forms of broadside array. Several types of curtain arrays were described by C. S. Franklin, two of which are shown in the figure. Figure 26 shows in somewhat greater detail the Sterba array, including feed lines, impedance matching devices and provisions for sleet melting. (See reference 5.) As will be noted by tracing the connections, it is possible without interrupting service to apply a 60-cycle power to this antenna for purposes of melting sleet, a provision found in earlier forms of linear conductor antennas (e.g., Fig. 22).

It has already been pointed out that the directional characteristic of a long straight wire may be used to produce antennas of marked directivity. Two forms are represented in Fig. 27. The one at the left, called a "rhombic" antenna (E. Bruce, reference 8), presents

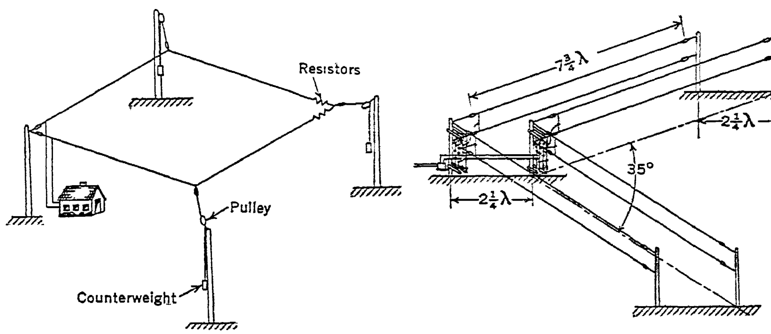


FIG. 27. Alternative Forms of Folded Wire Arrays

an impedance to the terminal equipment which has a constant resistive value, making it suitable over a wide frequency range (typically 2 to 1) provided that the accompanying shift of directivity is appropriate to the medium of propagation; in transatlantic work it is a fortunate circumstance that the higher frequencies used by day call for a more nearly horizontal ray than the lower frequencies used at night, making such an antenna suitable for a wide range of conditions. When the antenna is used for transmitting a considerable power has to be dissipated in the terminating resistance. For details of the resonant-V antenna shown at the right of Fig. 27, as well as several other interesting types, reference may be made to Carter, Hansell, and Lindenblad (reference 8). Long wires are also used in the vertical plane for directional effects, as in the "inverted-V antenna" shown in Fig. 28.

An interesting application of array principles to reception of transoceanic telephony is afforded by the Musa receiving antenna, the arrays of which are constructed with rhombic antennas as the units. (See reference 18.) Over routes such as this, radio waves of high frequency usually travel by several paths simultaneously and arrive at different angles above the horizontal. These components arrive with unrelated radio-frequency phases and even with differences of time delay which are significant in the audio range (of the order of a few tenths of a millisecond). The Musa antenna (Multiple-Unit-Steerable-Antenna) takes advantage of the spread in angle of arrival to separate the component waves, which may then be used singly, or in "diversity reception," or combined with audio delay correction, etc. Since the waves are more reliably distinguished by differences in vertical rather than azimuthal angle of arrival, the unit antennas are arranged along the great-circle path rather than in broadside array. One antenna used commercially has 16 unit rhombic antennas in an array 2 miles long. The combining of units is not done in the radio-frequency circuits, but it is accomplished at intermediate frequency through the medium of a common beating oscillator. A multiple system of phase shifters permits the separation and simultaneous reception of the different components provided that they are sufficiently different in angle of arrival.

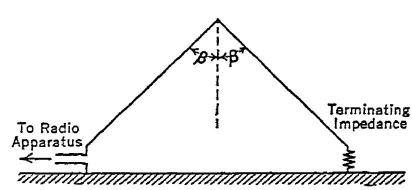


FIG. 28. Inverted-V Antenna

ANTENNAS FOR VERY HIGH AND ULTRA HIGH FREQUENCIES. The extensive application in point-to-point and mobile services of frequencies above 30 Mc has led to a diversity of antenna types. Figure 29 illustrates a few types useful for vertical electrical polarization and suitable for mounting on a pole. Except for one insulator which is specifically labeled, all lines shown represent conductive material, insulators being omitted in the interest of simple representation. The lower portion of (*a*) is the coaxial which feeds the upper $\lambda/2$ section at the middle in a series connection (the whole feed-line current flows into the antenna). The portion of large diameter does not touch the outer conductor of the feed line except at the top, an arrangement which tends to minimize waves standing on the supporting pole. In Fig. 29*b* there is a metallic conection between the top of the skirt and all adjacent parts; the feed is accomplished by bringing the inner conductor of the feed line through a hole in the outer one at a point inside the skirt which is protected from the weather. Figure 29*c* and *d* indicates two J-shaped antennas in which the radiating

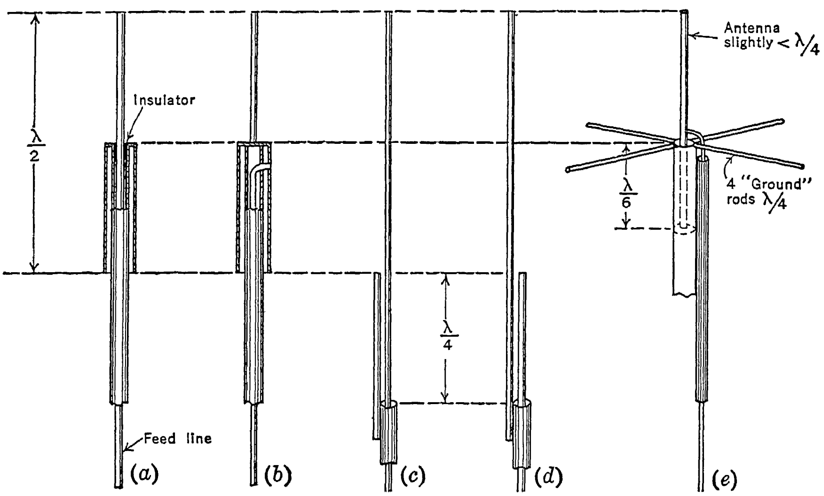


FIG. 29. Omnidirectional Antennas Using Vertical Polarization

section is the upper half-wave of one of the feed lines above the point where the other ends. (See reference 19.) Figure 29*e* shows a horizontal cross of four ground rods at the top of a large supporting cylinder, in the hollow end of which is mounted an inner conductor extending above it somewhat less than a quarter wavelength, the point of connection of the coaxial line being such as to match impedances. The section below this connection point provides a strong mechanical support for the radiating member above. (See reference 20.)

When horizontal polarization is used in ultra-high-frequency broadcasting there are several antenna types which may be considered. With this polarization it is easier than with vertical to increase the gain by "stacking." Several types are indicated in Fig. 30.

The "turnstile" antenna is shown in Fig. 30*a*. (See reference 21.) Essentially it has two half-wave horizontal radiating members crossed at 90° and phased in quadrature. It is fed by a system of transmission lines. When equal currents are used in the two radiators, the directional diagram in the horizontal plane is a circle deformed somewhat toward a square. The vertical separation between stacked elements is one-half wave. The turnstile antenna has been adapted for broad-band use by employment of large conductors and careful attention to detail. A cross-section of such an antenna on the Empire State Building is shown in Fig. 30*b*, where the cigar-shaped conductors and the adjacent central parts are surfaces of revolution about the lines *AC* and *BD*. Separate transmission lines are provided at *F* for each of the four radiators. (See references 21 and 24.)

Figure 30*c* is an "Alford loop," which is in the form of a horizontal square the length of whose edge is a matter of design, but which, for descriptive purposes, may be taken as of the order of one-third wavelength. Current is supplied as shown, the currents in the four radiating members being equal in magnitude and alike in phase as shown by the arrows in the diagram. In stacking a vertical spacing of one-half wave is used.

Figure 30*d* shows a circular antenna (see reference 22) which also is substantially a loop antenna. The two circular radiating conductors indicated are electrically broken at *B* by a parallel-plate condenser without loss of mechanical continuity and strength, the whole assembly being capable of support from point *A*. The lower circle is broken at *C*, from which point the system is fed in the manner of the "folded dipole," the "electrical length" of the circumference (taking account of the loading capacitance *B*) being one-half wave. Physically the circumference is less than this. This loop is attached to a vertical pole at *A* and is thus metallically grounded. The pole is inside the loop. The horizontal directional

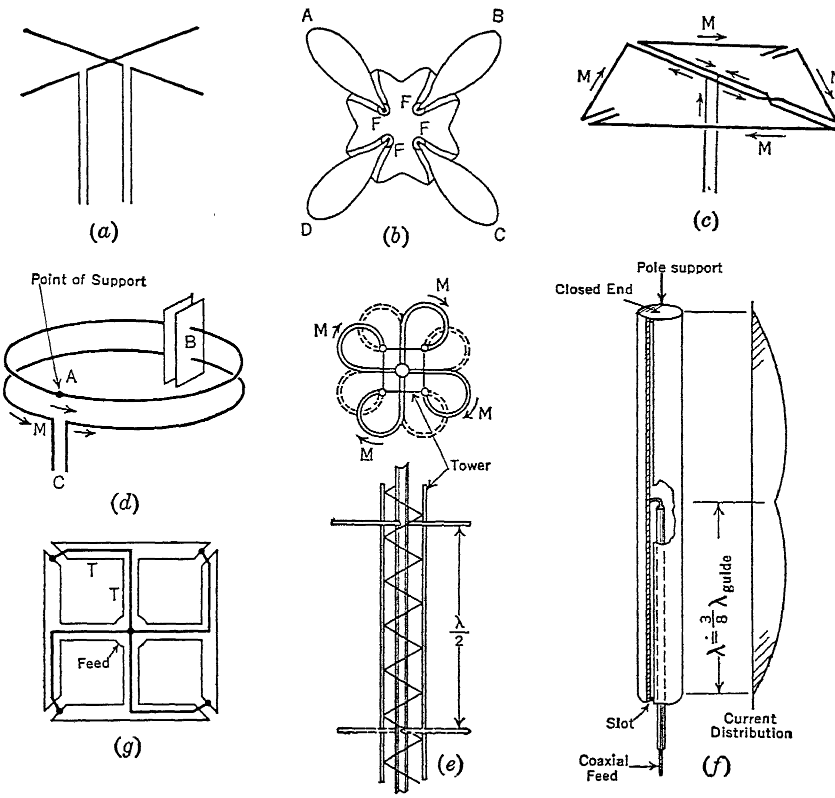


FIG. 30. Nondirectional Antennas Using Horizontal Polarization

pattern is elliptical, the maximum difference in field strength being somewhat less than 2 db. When these units are stacked the vertical spacing is one wavelength.

The "cloverleaf" antenna, due to P. H. Smith (reference 24) is shown in Fig. 30*e*. This consists of a slender tower (e.g., square) in the form of a conventional structural-steel lattice. Up the center is a conductor, which, together with the tower itself, forms a coaxial transmission system. The radiating "leaves" are attached as shown, forming a composite horizontal loop. The length of each of these conductors is about 0.4 λ. In stacking, a half-wave interval is used, and, because of the resulting phase reversal, a clockwise loop has counterclockwise loops immediately above and below it. Within a range from 88 to 108 Mc one antenna can be changed from one frequency to another by varying the vertical spacing between loops of one standard size. The horizontal diagram is substantially circular.

The "rocket" antenna, described by Andrew Alford and shown in Fig. 30*f*, is a vertical cylinder, metallically closed at both ends (in the form shown), but having an open slot along one element of the cylinder. It is fed as shown at the point where the cylinder is cut away by establishing a voltage across the slot. It may be thought of as a "lossy" wave guide supporting a transverse-electric mode, the critical frequency of which is (1) considerably less than that of the dominant mode of an unslotted cylinder owing to participation of the slot and of the space outside the cylinder in the propagation of the wave, and (2) somewhat less than the operating frequency. The metallic ends produce a standing-wave pattern which gives an approximation to uniform vertical distribution of radiating current over the outside, except near the ends. The antenna is in external effect somewhat like a vertical distribution of horizontal loops. In stacking the units are placed in close proximity. The field varies some 3 $^1/_2$ db in different azimuths. The diameter is somewhat less than one-half wavelength, and the driving-point impedance is tuned by folding the edges of the slot in. The "pylon" is a self-supporting antenna employing somewhat the same electrical principle.

A horizontal square loop employing an interesting coaxial feed system has been described (see reference 23) and is shown in Fig. 30*g*. The sides of the square are electrically 180° in length, and adjacent sides are fed at each corner in a pushpull manner from a vertical coaxial lying along the axis of symmetry, as shown. The "quarter-wave" sections T between the feed points and this coaxial also provide impedance transformation.

See reference 30 concerning some other antennas of interest in this frequency range.

MICROWAVE ANTENNAS. Practical microwave antennas employ many component devices in various combinations, but we can here recognize only general forms without elaboration. Brevity may lead to omission of types of greater future importance than those included, but this is unavoidable in a field so young and lusty.

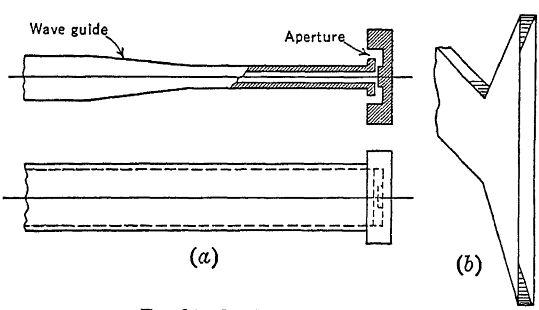


FIG. 31. Typical Point Sources

Devices used in the method of spherical optics are illustrated in Fig. 31, which depicts two point sources, and in Fig. 20, which shows corresponding methods of collimation (production of parallel rays). Figures 31*a* and 20*b* show a rear feed in which the energy comes through the wave guide from the left and is emitted from the two apertures *toward* the left. (See reference 26.) Where applicable, rear feeds for reflectors provide a desirable method of mechanical support without serious interference with the electrical functioning either of itself or of the reflector.

Conditions often dictate a "front" feed. Wave-guide apertures or horns are frequently used. One of the latter is illustrated in Fig. 31*b*, a sectoral horn in which the electric vector might with appropriate design be either vertical or horizontal. The horn shown could suitably be used with a paraboloidal reflector having an elliptical or rectangular shape, the greater dimension being the left-right one (note that this horn radiates a beam which is considerably wider horizontally than vertically if the aperture has a vertical dimension greater than a wavelength). A wave guide whose open end faces the reflector is often used (with or without a flare) particularly where the paraboloidal reflector subtends large angles both in elevation and azimuth.

When the feed is located in the path of the reflected wave, as in Fig. 20*b*, part of the energy will re-enter it and travel back along the transmission path into the transmitter. This may be great enough to cause trouble. One of the solutions for this difficulty is indicated in Fig. 20*a*, where the portion of the reflector which might cause a wave to be returned to the feed is omitted, and the feed, which is at the focus, is directed toward the active reflecting surface. A similar solution in which the path from feed to reflector is enclosed in a horn is indicated in Fig. 20*c*. Figure 20*d* shows how the spherically expanding wave from a point source can be collimated by a metal plate lens, such as has been described above. A lens in which zoning has been introduced to reduce lens thickness is indicated in Fig. 20*e*, while *f* suggests the advantage of a lens in shortening a highly directional horn.

Figure 19 illustrates some principles of design involving cylindrical optics. Figures 19*a* and 19*b* include line sources employing a linear array of dipoles and a sectoral parabola excited in the TE_{01} (rectangular) mode respectively. (Note that, according to the accepted convention, the designation TE_{10} of the most common mode in rectangular wave guides gives place to TE_{01} when the dimension parallel to the electric vector becomes greater than the other dimension of the cross-section.)

In some applications, such as air-borne radar, it may be necessary to have in azimuth a sharp concentration, and in elevation a wide spread of signal having intensity distributed according to some definite function of elevation angle. When this distribution gives uniform response along the ground the antenna is termed a *cosecant antenna.*

32. DIRECTION FINDING

Almost every type of directional antenna has, at some time or other, been used for direction finding, but there are a few types which are of especial importance due to their widespread use. (See reference 27.) Among those used in the lower frequencies are the loop in various forms and the Adcock antenna. Microwaves have their own distinctive methods, such as "lobing" and "scanning." The directivity may be in the receiver or in the transmitter, and in radar it is commonly used in both simultaneously.

The loop (see Fig. 1*k*) has great simplicity to recommend it but also these objectionable features to be avoided: it does not distinguish sense (e.g., east from west); it is sensitive to so-called antenna effect (response to electric vector independent of direction of arrival of the wave); and errors are found when the arriving waves have a downward direction.

Signals arriving in a horizontal direction induce emf's in a loop in proportion to its area and the magnetic field of the wave. At long waves most loops are small compared with a wavelength, and therefore they tend to pick up only small signals. This bona fide emf is therefore subject to interference from a spurious signal derived from the electric rather than the magnetic field (a more acceptable statement would be "from $\mathcal{E}$ rather than from

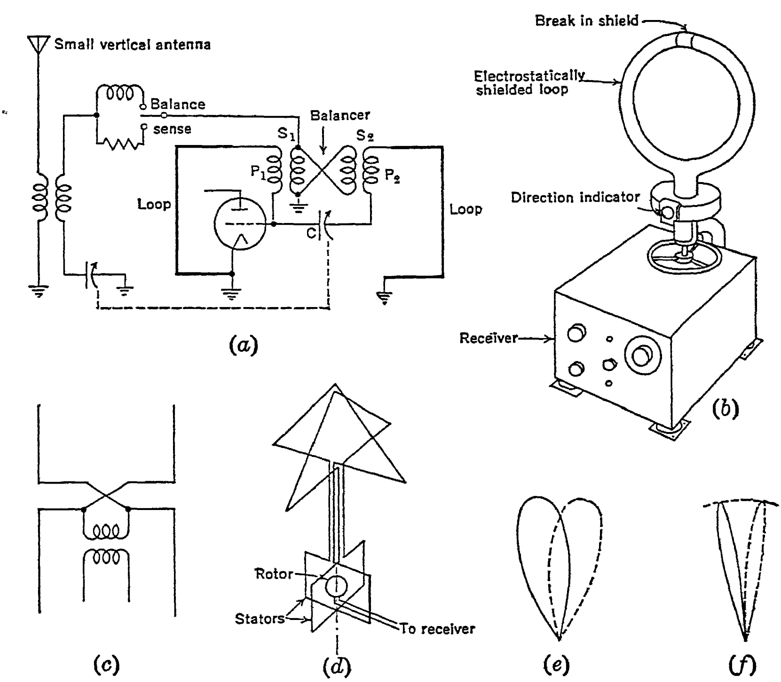


FIG. 32. Methods of Direction Finding

$\partial\mathcal{E}/\partial x$"). Although loops ideally possess symmetry which will suppress this component, as a practical matter it is likely to be an important source of error. Figure 32*a* (see reference 28) shows an application of a method for overcoming this difficulty by the addition of a signal from an auxiliary small vertical antenna to balance out this antenna effect, thus obtaining sharp nulls. The same circuit has provision by switching for altering the phase of the added signal by 90°, so that it aids one of the loop maxima and opposes the other, thus establishing the sense (east or west). A practical embodiment of this type of circuit is shown in Fig. 32*b* (see reference 28). It illustrates the important method of using a shielded loop to reduce the effect of local induction fields and to preserve electric symmetry.

The third kind of loop error occurs when the direction of arrival is not horizontal and there has been some rotation of the plane of polarization. It is really due to pick-up in the horizontal wires of the loop and can be combated by essentially getting rid of them. The Adcock antenna illustrated in Fig. 32*c* avoids this difficulty by eliminating the horizontal members and feeding the vertical ones by a shielded transmission line.

Figure 32*d* depicts the Bellini-Tosi loop method of direction finding in which the loops are fixed. By means of shielded transmission lines (usually horizontal rather than as shown in the figure) and a goniometer, the direction can be determined in an operating room somewhat removed from the antenna itself, which in this case may be a large one.

Figure 32*e* illustrates the principle of "lobing," in which by one of several devices the direction of beam can be alternated between the positions of the solid and the dotted lines. When the antenna is oriented so that the signals are equal, the intersection of the lobes indicates the direction. This is the method used in many radars.

Still another method is scanning, as indicated in Fig. 32*f*, where the beam, preferably very sharp, is swept periodically through a given range. The direction of maximum response gives the desired information.

33. MISCELLANEOUS

Antenna-testing methods are being studied by the Antenna Committee of the Institute of Radio Engineers, 1 East 79 Street, New York 21, N. Y. The report, when issued, will represent a much more complete document than that issued by the Committee in 1938. See reference 29. From the same source there is now available "Standards on Antennas, etc. Definitions of Terms," price 75 cents.

For information on *transmission lines* see reference 25.

For information on *antennas for aircraft* see reference 31.

BIBLIOGRAPHY

GENERAL

Federal Communications Commission, "Standards of Good Engineering Practice Concerning Standard Broadcast Stations." Effective Aug. 1, 1939, revised to June 1, 1944. For sale by the Superintendent of Documents, Washington 25, D. C.

Ladner, A. W., and C. R. Stoner, *Short Wave Wireless Communications*. John Wiley, New York (1933).

McIlwain, K., and J. G. Brainerd, *High-frequency Alternating Currents*. John Wiley, New York (1939).

Schelkunoff, S. A., *Electromagnetic Waves*. Van Nostrand, New York (1943).

Terman, F. E., *Radio Frequency Engineering*, Chapter XIV. McGraw-Hill, New York (1937).

1. Schelkunoff, S. A., *Electromagnetic Waves*, Chapter XI. Van Nostrand, New York (1943).
   Theory of Antennas of Arbitrary Size and Shape, *Proc. I.R.E.*, Vol. 29, No. 9, pp. 493–521 (September 1941).
2. Ballantine, Stuart, On the Radiation Resistance of a Simple Vertical Antenna at Wave Lengths below the Fundamental, and, On the Optimum Transmitting Wave Length for a Vertical Antenna over Perfect Earth, *Proc. I.R.E.*, Vol. 12, No. 6, pp. 823 and 833 (December 1924).
3. Foster, Ronald M., Directive Diagrams of Antenna Arrays, *B.S.T.J.*, Vol. V, No. 2, p. 292 (April 1926).
   Southworth, G. C., Certain Factors Affecting the Gain of Directive Antennas, *Proc. I.R.E.*, Vol. 18, No. 9, p. 1502 (September 1930).
   Sterba, E. J., Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. I.R.E.*, Vol. 19, No. 7, p. 1184 (July 1931).
4. Friis, H. T., and W. D. Lewis, *Radar Antennas*. *B. S. T. J.*, Vol. 26, No. 2, 219–317 (April 1947).
5. Sterba, E. J., Theoretical and Practical Aspects of Directional Transmitting Systems, *Proc. I.R.E.*, Vol. 19, No. 7, pp. 1194–1196 (July 1931).
   On calculations of characteristics of arrays, etc., see: A. A. Pistolkors, The Radiation Resistance of Beam Antennas, *Proc. I.R.E.*, Vol. 17, No. 3, p. 462 (March 1929); K. Bechmann, On the Calculation of Radiation Resistance of Antenna and Antenna Combinations, *Proc. I.R.E.*, Vol. 19, No. 8, p. 1471 (August 1931); P. S. Carter, Circuit Relations in Radiating Systems and Applications to Antenna Problems, *Proc. I.R.E.*, Vol. 20, No. 6, p. 1004 (June 1932); and Irving Wolff, Determination of the Radiating System Which Will Produce a Specified Directional Characteristic, *Proc. I.R.E.*, Vol. 25, No. 5, p. 630 (May 1937).
6. Duttera, W. S., Some Factors in the Design of Directive Broadcast Antenna Systems, *R.C.A. Rev.*, Vol. II, No. 1, pp. 81–93 (July 1937).
   Brown, G. S., Directional Antennas, *Proc. I.R.E.*, Vol. 25, No. 1, pp. 78–145 (January 1937).
   Guy, Raymond F., The New FCC Regulations, *Electronics*, August 1939, p. 11.
7. Morrison, J. F., A Simple Method for Observing Current Amplitude and Phase Relations in Antenna Arrays, *Proc. I.R.E.*, Vol. 25, No. 10, p. 1310 (October 1937).
   Kear, F. G., Maintaining the Directivity of Antenna Arrays, *Proc. I.R.E.*, Vol. 22, No. 7, p. 847 (July 1934).
8. Bruce, E., Developments in Short-Wave Directive Antennas, *Proc. I.R.E.*, Vol. 19, No. 8, p. 1406 (August 1931).
   Carter, P. S., C. W. Hansell, and N. E. Lindenblad, Development of Directive Transmitting Antennas by R.C.A. Communications, Inc., *Proc. I.R.E.*, Vol. 19, No. 10, p. 1773 (October 1931).
   Bruce, E., A. C. Beck, and L. R. Lowry, Horizontal Rhombic Antennas, *Proc. I.R.E.*, Vol. 23, No. 1, p. 24 (January 1935).
   Harper, A. E., *Rhombic Antenna Design*. Van Nostrand, New York (1941).

9. Potter, R. K., and H. T. Friis, Some Effects of Topography and Ground on Short-Wave Reception, *Proc. I.R.E.*, Vol. 20, No. 4 (April 1932).
10. Friis, H. T., and W. D. Lewis, *Radar Antennas*, Section 3.3, of reference 4, above.
11. Cutler, C. C., Parabolic Antenna Design for Microwaves. *Proc. I.R.E.*, Vol. 35, No. 11, 1284 (November 1947).
12. Kock, W. E., Metal Lens Antennas, *Proc. I.R.E.*, Vol. 34, No. 11, p. 828 (November 1946).
Kock, W. E., Metallic Delay Lenses, *B.S.T.J.*, Vol. 27, No. 1, 58 (January 1948).
13. Electromagnetic Horns, see: Barrow, W. L., and F. D. Lewis, The Sectoral Electromagnetic Horn, p. 41, and W. L. Barrow and L. J. Chu, Theory of the Electromagnetic Horn, p. 51, *Proc. I.R.E.*, Vol. 27, No. 1 (January 1939); G. C. Southworth and A. P. King, Metal Horns as Directive Receivers of Ultra-short Waves, *Proc. I.R.E.*, Vol. 27, No. 2, p. 95 (February 1939); S. A. Schelkunoff, *Electromagnetic Waves*, Section 9.23 (Electric Horns), p. 360, Van Nostrand, New York (1943); A. P. King, *The Radiation Characteristics of Conical Horns* (to be published).
14. Alexanderson, E. F. W., Transoceanic Radio Communication, *Trans. A.I.E.E.*, Vol. 38, 1269–1285 (1919).
Alexanderson, E. F. W., Reoch, and Taylor, The Electrical Plant of Transocean Radio Telegraphy, *Trans. A.I.E.E.*, Vol. 42, 707–717 (1923).
Lindenblad, N. and W. W. Brown, Main Considerations in Antenna Design, *Proc. I.R.E.*, Vol. 14, 291 (June 1926).
15. Feldman, C. B., The Optical Behavior of the Ground for Short Radio Waves, *Proc. I.R.E.*, Vol. 21, No. 6, p. 764 (June 1933).
Beverage, Rice, and Kellogg, The Wave Antenna—A New Type of Highly Directive Antenna, *Trans. A.I.E.E.*, Vol. 42, 215 (1923).
16. Gihring, H. E., and G. H. Brown, General Considerations of Tower Antennas for Broadcast Use, *Proc. I.R.E.*, Vol. 23, No. 4, p. 311 (April 1935).
Morrison, J. F., and P. H. Smith, The Shunt-Excited Antenna, *Proc. I.R.E.*, Vol. 25, No. 6, p. 673 (June 1937).
17. Rettenmeyer, F. X., Radio Frequency Distributing Systems, *Proc. I.R.E.*, Vol. 23, No. 11, p. 1286 (November 1935).
Wheeler, H. A., and V. E. Whitman, The Design of Doublet Antenna Systems, *Proc. I.R.E.*, Vol. 24, No. 10, p. 1257 (October 1936).
Carlson, W. L., and V. D. Landon, A New Antenna Kit Design, *R.C.A. Rev.*, Vol. II, No. 1, p. 60 (July 1937).
Landon, V. D., and J. D. Reid, A New Antenna System for Noise Reduction, *Proc. I.R.E.*, Vol. 27, No. 3, p. 188 (March 1939).
18. Friis, H. T., C. B. Feldman, and W. M. Sharpless, The Determination of the Direction of Arrival of Short Radio Waves, *Proc. I.R.E.*, Vol. 22, No. 1, p. 47 (January 1934).
Friis, H. T., and C. B. Feldman, A Multiple Unit Steerable Antenna for Short-Wave Reception, *Proc. I.R.E.*, Vol. 25, No. 7, p. 841 (July 1937).
Polkinghorn, F. A., A Single Sideband Musa Receiving System for Commercial Operation on Transatlantic Radio Telephone Circuits, *B.S.T.J.*, Vol. 19, No. 2, p. 306 (April 1940).
19. Tinus, W. C., Ultra-high Frequency Antenna Terminations Using Concentric Lines, *Electronics*, Vol. 8, p. 239 (August 1935).
20. Brown, G. H., and J. Epstein, An Ultra-high-frequency Antenna of Simple Design, *Communications*, Vol. 20, No. 7, p. 3 (July 1940).
21. Brown, G. H., and J. Epstein, A Pretuned Turnstile Antenna, *Electronics*, Vol. 18, No. 6, p. 102 (June 1945). (Gives bibliography.)
[Lindenblad, N. E., Television Transmitting Antenna for Empire State Building, *R.C.A. Rev.*, Vol. 3, No. 4, p. 387 (April 1939).
Carter, P. S., Simple Television Antennas, *R.C.A. Rev.*, Vol. 4, No. 2, p. 168 (October 1939).
22. Scheldorf, M. W., FM Circular Antenna, *Gen. Elec. Rev.*, Vol. 46, No. 3, p. 163 (March 1943).
23. Kandoian, A. G., Three New Antenna Types and Their Applications, *I.R.E.*, Waves and Electrons, Vol. 1, No. 2, p. 70W (February 1946).
24. A group of short discussions on antennas for FM broadcasting, i.e., circular, broadcast loops, super-turnstile, and cloverleaf antennas, Communications, Vol. 26, pp. 40, 58 (April 1946).
25. For information on transmission lines see Section 10 in this volume (Wire Transmission Lines); also see W. S. Duttera, New Coaxial Conductor at WTAM, *Electronics*, March 1939, p. 30. J. B. Epperson, Installation of Coaxial Transmission Lines, Part I, *Electronics*, July 1939, p. 30, August 1939, p. 31 (Part II; E. J. Sterba and C. B. Feldman, Transmission Lines for Short-wave Radio Systems, *Proc. I.R.E.*, Vol. 20, No. 7, p. 1163 (July 1932); W. L. Everitt and J. F. Byrne, Single-wire Transmission Lines for Short-Wave Antennas, *Proc. I.R.E.*, Vol. 17, No. 10, p. 1840 (October 1929).
26. Cutler, C. C., Design of Parabolic Antennas. See reference 11 above.
27. Keen, R., *Wireless Direction Finding*, 3d Ed. Iliffe & Sons, London (1938).
28. Martin, H. B., Small Vessel Direction Finding, *R.C.A. Rev.*, Vol. II, No. 1, p. 69 (July 1937).
29. Report of the Antenna Committee entitled: "Standards on Antennas," to be published by the Institute of Radio Engineers, 1 E 79 Street, New York 21, N. Y.
Cutler, C. C., A. P. King, and W. E. Kock, Microwave Antenna Measurements. *Proc. I.R.E.*, Vol. 35, No. 12, 1462 (December 1947).
30. Hidetsugu Yagi, Beam Transmission of Ultra Short Waves, *Proc. I.R.E.*, Vol. 16, No. 6, p. 715 (June 1928).
Kraus, John D., The Corner-Reflector Antenna, *Proc. I.R.E.*, Vol. 28, No. 11, p. 513 (November 1940).
31. Antennas for Aircraft: P. C. Sandretto, *Principles of Aeronautical Engineering*, McGraw-Hill, New York (1942); F. D. Bennett, P. D. Coleman, and A. S. Meier, The Design of Broad-Band Aircraft-Antenna Systems, *Proc. I.R.E.*, Vol. 33, No. 10, p. 671 (October 1945); Flt. Lt. C. B. Bovill, *J. A.I.E.E.* (London), Vol. 92, Part III, p. 105 (June 1945); G. L. Haller, Aircraft Antennas, *Proc. I.R.E.*, Vol. 30, No. 8, p. 357 (August 1942).

SECTION 7

VACUUM-TUBE CIRCUIT ELEMENTS

VACUUM-TUBE CIRCUIT ELEMENTS

AMPLIFIERS

By Loy E. Barton

An amplifier is a device for increasing the energy associated with any phenomenon without appreciably altering its quality.

Amplifiers used in communication circuits almost invariably employ thermionic vacuum tubes as the amplifying elements. The vacuum tube is practically a pure resistance device at low frequencies. As the frequency increases to a point where the interelectrode capacity impedance becomes appreciable with respect to the internal resistance of the tube a phase shift is introduced which alters gain or amplification.

At still higher frequencies the time required for the electrons to reach the plate is appreciable with respect to a quarter cycle. When this transit time is about 20 to 30 electrical degrees the amplifier characteristics are modified considerably. These amplifiers still fall in one of the three basic classes of amplifiers.

A class A amplifier is an amplifier in which the grid bias voltage permits a steady plate current flow of such a value that the plate current varies directly as the grid voltage for the complete cycle of 360 electrical degrees. The resulting output voltage for an ideal class A amplifier is an exact reproduction of the grid voltage.

The characteristics of the class A amplifier are low power output with a theoretical maximum plate power efficiency of 50 per cent and an operating efficiency of about 30 per cent or less at full power outputs. The plate dissipation is maximum at zero output, and the plate circuit output network may be tuned or untuned for an undistorted output. The average value of plate current does not change during the cycle so that the input plate power is constant.

A class B amplifier is an amplifier in which the grid voltage permits essentially zero plate current with no signal applied to the grid and the plate current is proportional to the grid swing when the grid swings in a positive direction from the bias point so that plate current flows for approximately 180 electrical degrees.

The characteristics of the class B amplifier are comparatively high power output with a theoretical maximum plate efficiency of 78.5 per cent; in practice the efficiency approaches 65 per cent at full output. The plate dissipation is a minimum and is comparatively low at zero signal, increasing rapidly to approximately constant value at about 25 per cent full output. However, the power input to the plate increases, with signal and power output, until the peak output is reached. Therefore, the plate current is a variable and the plate supply voltage should have good regulation. The class B amplifier may be used as a single-tube tuned-plate circuit amplifier or the plate circuit may be untuned provided two tubes are used in a pushpull manner with appropriate input and output transformers.

A class B_1 amplifier is an amplifier biased and operated as a normal class B amplifier except that the grid swing does not go into the positive region.

In general such an amplifier uses low-mu tubes in order that high plate currents may be reached without driving the grid into the grid-current region. The efficiency of this amplifier is lower than that of an amplifier in which the grid is driven to plate-current saturation, and the power output is lower. However, the grids take no power so that input circuit losses and distortion may be low.

A class AB_1 amplifier is an amplifier so biased that the pushpull tubes act as a class A amplifier for low grid swings and go into class B_1 operation at higher grid swings. The grids are not driven positive.

The only advantage of this type of operation is that somewhat higher outputs may be obtained for a given plate dissipation than can be obtained for class A operation and a cathode resistor can be used for self-bias. However, distortion is comparatively high.

A class AB_2 amplifier is similar to the class AB_1 amplifier except that the grids are driven into the positive region for higher outputs.

The characteristics of the class AB_2 amplifier are much the same as those of the class AB_1 except that the power output is higher, and the grid driving problem is about equal to that of the class B amplifier.

A class C amplifier is an amplifier in which the grid bias voltage is appreciably higher than the bias required for plate-current cutoff, and the plate current flows for a period less than 180 electrical degrees during the half cycle when the grid swing is positive with respect to the bias voltage. The grid swing is usually to the point of plate-current saturation, in which case the rms plate current is proportional to plate voltage and is not proportional to the grid voltage.

The characteristics of the class C amplifier are high power output and an average plate efficiency in practice of 70 to 75 per cent, which may reach 85 to 90 per cent under special conditions. (See Fig. 35, p. 7-27, and discussion.) The theoretical maximum efficiency for the class C amplifier is 100 per cent.

1. CLASS A AMPLIFIERS

GENERAL USE. The class A amplifier is used for audio-frequency and radio-frequency voltage amplification, principally because the output voltage is a direct function of the input voltage. The amplifying action of the vacuum tube follows from the equations developed for its plate current in Section 5. If all the terms but the first in eq. (17), p. 5-43, are negligible, the voltage drop due to the plate current of the vacuum tube flowing through an external impedance is

$$e_z = zi_{pa} = \mu e_g \left[\times \frac{z}{r_p + z} \right] \tag{1}$$

or for a particular component of periodicity ω_m

$$E_z = \frac{\mu E_g z_m}{r_p + z_m} \tag{1a}$$

This voltage is a *magnified* replica of the grid voltage and is said to be amplified.

The voltage amplification is the ratio of the change in voltage across the external load to the change in input voltage, or

$$\text{V.A.} = \frac{\mu\sqrt{r^2 + x_m^2}}{\sqrt{(r + r_p)^2 + x_m^2}} = \frac{\mu z_m}{z'_m} \tag{2}$$

if there is no drop in the external grid circuit. Two special cases are of interest: First, when the load is a pure resistance; V.A. $= \mu r/(r_p + r)$ which will equal 0.9 μ when $r = 9r_p$. Second, when the load is a pure reactance, V.A. $= \mu x_m/\sqrt{r_p^2 + x_m^2}$, which will approximate 0.9 μ when $x_m = 2r_p$.

The power amplification is the ratio of the power delivered by the output circuit to the power supplied to the input circuit, or

$$\text{P.A.} = \frac{\mu^2 r z''^2}{[(r_p + r)^2 + x_m^2](r_i + r_e)} \tag{3}$$

where z'' is the total impedance in the grid circuit.

For small power outputs, the power amplification is very high and is usually limited only by the input circuit losses.

Cascade Amplifiers. The limitations of voltage and power amplification obtainable with one tube make it frequently desirable to use two or more tubes in cascade by coupling the plate circuit of the first tube to the grid circuit of the next tube, etc. Each tube with its associated circuits is called a *stage* of the amplifier, and the whole is termed a multistage or cascade amplifier. Cascade amplifiers are usually classified by the method of coupling used.

The amplification per stage is defined as the ratio of the grid voltage of one tube to the grid voltage of the preceding tube. This is frequently expressed in decibels.

For the resistance-coupled amplifier shown in Fig. 1, it is

$$\text{V.A.} = \frac{\mu}{\dfrac{r_p r_{i2}}{z_{i2}^2} + \dfrac{r_p}{r} + 1 - j\dfrac{r_p x_{i2}}{z_{i2}^2}} \tag{4}$$

For low frequencies, and when the grid biasing voltage is always negative, the amplification limit per stage of this type of amplifier is μ.

$$\text{V.A.} = \frac{\mu r}{r_p + r} \tag{5}$$

A direct resistance coupled amplifier of the type shown in Fig. 1 is used when it is necessary that a d-c voltage be amplified. It will be noted that the required B voltage increases with each stage and that independent filament supplies are needed for direct-heated tubes. In the case of indirectly heated cathodes a separate heater supply is usually needed because of the excessive heater cathode potential.

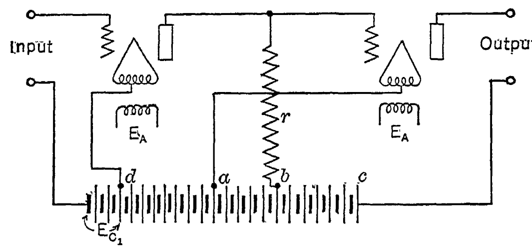


FIG. 1. Direct Resistance Coupled Amplifier, Common B and C Batteries, Separate A Batteries

The amplification per stage of the resistance-capacitance coupled amplifier shown in Fig. 2 is

$$\text{V.A.} = \frac{\mu}{\dfrac{z_c + r_p}{z_{i2}} + \dfrac{z_c + r_p}{r_K} + \dfrac{r_p z_c}{r r_{i2}} + \dfrac{r_p z_c}{r r_K} + \dfrac{r_p}{r} + 1} \tag{6}$$

in which z_c is the reactance of the coupling condenser C. A reactance may be substituted for either r or r_K, or both, in which case these constants are changed from r to z in eq. (6).

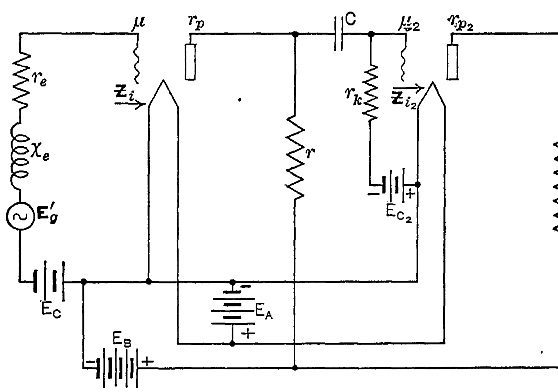


FIG. 2. Resistance-capacitance Coupling

If the input impedance to the second tube is very high and the impedance of the coupling condenser negligible eq. (6) reduces to

$$\text{V.A.} = \frac{\mu}{\dfrac{r_p}{r_K} + \dfrac{r_p}{r} + 1} \tag{6a}$$

from which it is seen that the higher the resistance of the parallel combination of r and r_K the greater the realized amplification will be. The limiting value of r_K is usually 1 megohm or less, so that it is of little value to make r greater than 75,000 to 100,000 ohms.

The plate voltage supply must be great enough to operate the tube at a desirable point and to supply the drop in the load resistor in order to obtain maximum voltage in the succeeding tube. A lower plate supply voltage may be used with a corresponding reduction of bias if maximum voltage output is not required. This will have little effect on the amount of amplification obtained.

Cascade operation of two or more resistance-capacitance coupled audio amplifier stages is subject to a low-frequency oscillation or "motor-boating," resulting from a feedback of low frequencies due to the common a-c impedance of the plate supply at frequencies so low that the filter condensers are not effective. To correct the motor-boating difficulty, it is usually necessary to limit the low-frequency response or isolate one or more of the resistance coupled stages. The low-frequency response is most easily reduced by decreasing the size of the coupling condenser (C in Fig. 2).

A transformer-coupled amplifier is shown in Fig. 3 and its equivalent circuit in Fig. 4. In the equivalent circuit, account is taken of the distributed capacitances (C_2 and C_5)

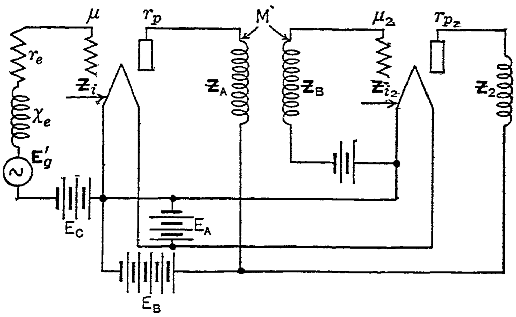


FIG. 3. Transformer Coupled Amplifier

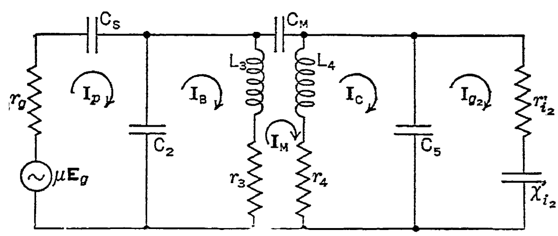


FIG. 4. Equivalent Circuit of the Transformer Coupled Amplifier of Fig. 3

of the windings and of the distributed capacitance (C_M) between windings. The transformer windings are assumed to be poled so that the effect of this latter is a minimum. The voltage amplification of each stage is then

$$\text{V.A.} = \frac{\mu z_M z_{i2}}{\left[z_i + \left(\dfrac{z_i + jx_2}{jx_2}\right) z_x\right]\left[z_{i2} + \left(\dfrac{z_{i2} + jx_5}{jx_5}\right) z_y\right] + \left(\dfrac{z_{i2}}{x_5} + j\right)\left(\dfrac{z_1}{x_2} + j\right) z^2_M} \quad (7)$$

where

$$z_M = \frac{z_3 z_4 + \omega^2 M^2 + \dfrac{M}{C_M}}{z_3 + z_4 - j2\omega M - \dfrac{j}{\omega C_M}}$$

$$z_x = \frac{z_3 z_4 + \omega^2 M^2 - j\dfrac{z_3}{\omega C_M}}{z_3 + z_4 - j2\omega M - \dfrac{j}{\omega C_M}}$$

and

$$z_y = \frac{z_3 z_4 + \omega^2 M^2 - j\dfrac{z_4}{\omega C_M}}{z_3 + z_4 - j2\omega M - \dfrac{j}{\omega C_M}}$$

Note that if the incidental resistances of the transformer windings are neglected z_M, z_x, and z_y are all pure imaginaries.

The maximum value of V.A. with respect to x_5 occurs under two conditions: one value (infinity) will give maximum value of V.A. for all frequencies; the other value, dependent in a complicated manner on the mesh parameters, will give a maximum for only one particular frequency. The same result is found when x_2 is considered the variable. These correspond to the cases of tuned and untuned transformers. (See p. 6-08.)

When substantially equal amplification is required over a broad frequency range the distributed capacitance should be made as small as possible. Assuming x_2 and x_5 infinite and C_M zero, eq. (7) reduces to

$$\text{V.A.} = \frac{\mu\omega M z_{i2}}{(z_1 + z_3)(z_{i2} + z_4) + \omega^2 M^2} \tag{7a}$$

By tuning primary and secondary circuits as shown in Fig. 5, the maximum amplification (at one frequency only) is

$$\text{V.A.} = \frac{\mu z_{i2}}{2\sqrt{(r_p + r_3)(r_{i2} + r_4)}} \tag{7b}$$

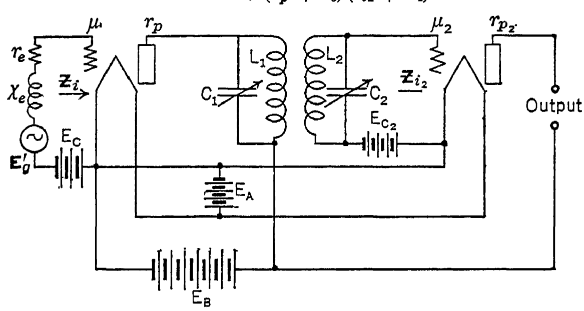


FIG. 5. Tuned Transformer Coupled Amplifier

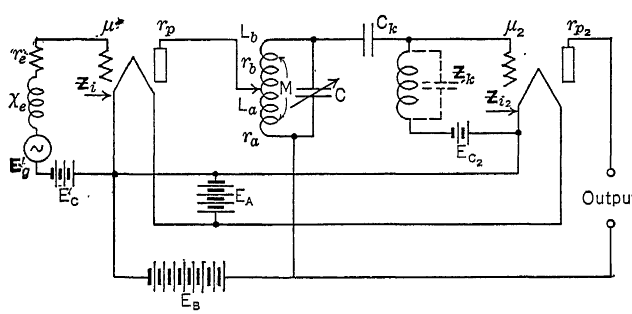


FIG. 6. Tuned Coupling. The tuned circuit is L_a–L_b–C. The coil Z_k and the capacitor C_k are inserted to get proper bias on the grid tube 2.

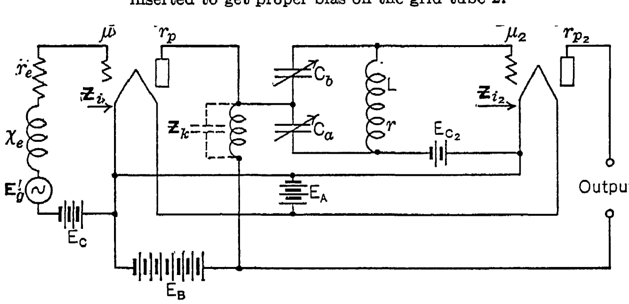


FIG. 7. Tuned Coupling. The tuned circuit is C_a–C_b–L. The coil Z_k is inserted to get proper potential on the plate of the first tube.

Other forms of tuned amplifiers are shown in Figs. 6 and 7. For the resonant frequency the amplification of the circuit of Fig. 7 is

$$\text{V.A.} = \frac{\mu}{\left(\frac{r}{z} - \frac{j\omega L}{z_{i2}}\right)(1 + j\omega C_a r_p) + \frac{j}{\omega C_a z} + \frac{j}{\omega C_a z_{i2}}} \tag{8}$$

where $z = r + j\omega L$.

Performance. Certain tube characteristics are supplied by the tube manufacturers (see Section 4) which are used to predict the performance of tubes as class A amplifiers; representative curves and tube constants will be given herein with sample calculations of performance. A typical general-purpose triode is the 56-type tube whose plate characteristics are given in Fig. 8. It will be noted that each curve of plate current against plate voltage is drawn for a given grid or bias voltage denoted as E_c. From these curves, the various constants of the tube may be calculated as explained in the section on vacuum tubes. The constants for the 56, as calculated from Fig. 8, at approximately 238 volts at the plate, 13.5 volts bias, and 5 ma plate current, are plate resistance (r_p) equals 10,000 ohms, amplification factor (μ) equals 15, and the grid plate transconductance (s_M) equals 1500 micromhos.

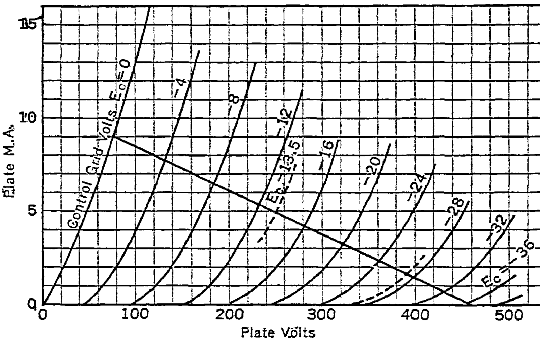


FIG. 8. Type 56, Plate Characteristics

INPUT CIRCUIT CALCULATIONS. A typical circuit for the 56 tube as a transformer-coupled audio amplifier is shown in Fig. 9 and its approximate equivalent in Fig. 10. If the tube used in Fig. 9 is the 56 type with characteristics as shown in Fig. 8, and its various constants are as indicated above, the constants of the circuit and performance of the tube may be calculated as follows:

The self bias resistor, r_1, is calculated for a 13.5-volt operating grid bias voltage with 5 ma plate current, in which case, according to the curves in Fig. 8, the voltage across the terminals marked minus and plus for plate supply should be 250 volts and the actual plate voltage will be about 236.5. The value of the resistor, r_1, is $13.5/0.005 = 2700$ ohms.

The by-pass condenser, C_1, must effectively by-pass r_1 at the lowest audio frequency desired so that the pulsating plate current through r_1 will not generate an audio voltage appreciable compared to that applied to the grid. The voltage across r_1 is out of phase with the input voltage so that the low-frequency gain may be lower than the gain at other

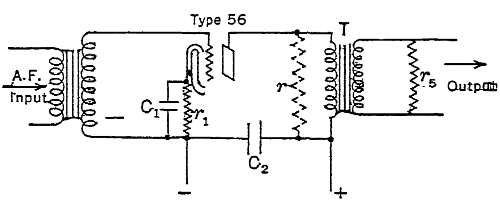


FIG. 9. Transformer-coupled Class A Amplifier

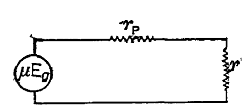


FIG. 10. Equivalent Circuit for Fig. 9

frequencies if the impedance of C_1 is large. The capacitance, C_2, across the plate supply may be the plate-supply filter condenser but, in any case, must have sufficiently low impedance at the lowest desired frequency to by-pass the plate supply effectively unless the a-c impedance of the plate supply is very low.

OUTPUT CALCULATIONS. The dotted resistance r in Fig. 9 is the equivalent impedance (as nearly unity power factor as possible) of the primary of the transformer, T, with the load resistance, r_5, transferred from the secondary into the primary. It will be noted that r is in series with the plate of the tube and may be represented by a load line drawn through the operating point at 5 ma plate current and 13.5 volts bias in Fig. 8.

The proper value for this resistance is obtained by drawing a line through the operating point and the point on the zero grid voltage curve representing the maximum desired value of the operating plate current (in this case 9 ma, since experience has shown that an 80 per cent change in plate current is a reasonable value to assume). A check along the load line as drawn indicates that, if the grid swings in a negative direction to double the grid voltage of 13.5 volts, the decrease in plate current is only 3.2 ma, approximately.

This load line represents approximately 43,000 ohms and is r as shown in Fig. 9. The approximate second harmonic under these conditions may be calculated by the following formula:

$$\left.\begin{aligned}\frac{I_{b\,\max} + I_{b\,\min} - 2I_b}{2(I_{b\,\max} - I_{b\,\min})} &= \text{per cent second harmonic approximately}\\ &= \text{5.5 per cent for 80 per cent increase in plate current as assumed above}\end{aligned}\right\} \quad (9)$$

in which I_b is the "no-signal plate current," $I_{b\,\max}$ is the peak plate current, and $I_{b\,\min}$ is the minimum plate current.

Note that if the maximum current change is equal on each side of the operating value the second harmonic is zero, since $(I_{b\,\max} - I_b) - (I_b - I_{b\,\min}) = 0$. If intermediate points are considered and it is found that equal grid-voltage increments do not cause equal plate-current increments throughout the range, higher harmonics will be present. (See p. 5-47.)

The average power output may be calculated by the following formula, which is readily derived from peak values of a-c voltages and currents.

$$\frac{(I_{b\,\max} - I_{b\,\min})(E_{b\,\max} - E_{b\,\min})}{8} = \left.\begin{array}{l}\text{power output or 0.28 watt approximately,}\\ \text{for full grid swing in the above case}\end{array}\right\} \quad (10)$$

in which $E_{b\,\max}$ is the maximum plate voltage and $E_{b\,\min}$ is the minimum plate voltage.

The theoretical maximum $I_{b\,\max}$ for the class A amplifier is $2I_b$, and the minimum $I_{b\,\min}$ is zero. The corresponding theoretical maximum $E_{b\,\max}$ is $2E_b$, and the minimum $E_{b\,\min}$ is zero.

The plate power input is expressed by the formula E_bI_b = watts input. Therefore, the plate efficiency at maximum output is

$$\text{Pl. eff.} = \frac{(I_{b\,\max} - I_{b\,\min})(E_{b\,\max} - E_{b\,\min})}{8E_bI_b} = 23 \text{ per cent} \quad (11)$$

The load resistance is

$$r = \frac{(E_{b\,\max} - E_{b\,\min})}{(I_{b\,\max} - I_{b\,\min})} = 43{,}000 \text{ ohms approximately} \quad (12)$$

The total plate voltage swing is approximately 314 volts, and the total grid swing is 27 volts, which is

$$\frac{(E_{b\,\max} - E_{b\,\min})}{(E_{c\,\max} - E_{c\,\min})} = \frac{314}{27} = 11.6 \text{ approximate actual amplification by the tube} \quad (13)$$

In the above circuit the transformer T may be designed as a voltage coupling transformer to supply a secondary load as high as 200,000 ohms or more, an output impedance which is not particularly difficult to obtain especially if a choke voltage feed is used as shown in Fig. 11, so that the d-c plate current does not go through the transformer pri-

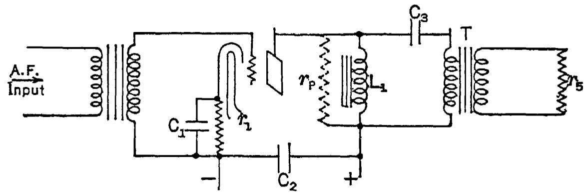


FIG. 11. Shunt Transformer-coupled Amplifier

mary. The turn ratio would be the square root of the impedance ratio or a turn ratio of the primary to secondary of at least 2.2. The total amplification of this amplifier stage including the transformer is 11.6 × 2.2 = 25.5, approximately.

The above condition is also approximately the condition for maximum power output. If the calculated distortion is too high, the grid swing may be reduced with a corresponding reduction of power output. In actual practice, the grid swing is usually very small if the amplifier is used as a voltage amplifier so that maximum voltage output is not obtained but the voltage amplification is the same; in such a case the second harmonic may be negligible (far less than 1/2 per cent).

PERFORMANCE CALCULATIONS FROM TUBE CONSTANTS. Referring to the equivalent circuit shown in Fig. 10, the voltage amplification of the combination of Fig. 9 and Fig. 8 may be calculated from the constants of the tube by means of the following formulas:

$$\text{V.A.} = \frac{\mu E_g \left(\dfrac{r}{r_p + r}\right)}{E_g} = \mu\left(\frac{r}{r_p + r}\right)$$
$$= 12.2 \text{ gain without transformer} \tag{14}$$

in which E_g is the rms grid voltage and the other constants are as defined above.

The power output for full grid swing of 13.5 volts peak or an rms value of 9.6 volts is

$$\text{P. out} = \frac{\left[E_g\mu\left(\dfrac{r}{r_p - r}\right)\right]^2}{r_p} = 0.27 \text{ watt} \tag{15}$$

It may be seen that, by using the constants of the tube in the equivalent circuit of Fig. 10, the performance calculations do not differ materially from the performance calculated from the curves in Fig. 8. The small discrepancy may be accounted for by the fact that the constants of μ and r_p used in eqs. (14) and (15) are obtained by using smaller increments of voltage and current changes than are used in eqs. (10) and (13); in other words, the effect of distortion is neglected.

It will be seen that, if the circuit shown in Fig. 9 is used, the only difference between a voltage and power amplifier is the design of the transformer, T, which in a voltage amplifier works into a high resistance and in a power amplifier works into the desired load, the turn ratio being chosen to present the same equivalent primary impedance.

If in the resistance coupled amplifier the load resistance is 43,000 ohms, the gain is the same as calculated for the above case if $r_K \gg r$. The same load line as shown in Fig. 8 may be used provided the plate supply voltage is equal to the value indicated by the intersection of the load line and the zero plate-current axis. This plate voltage is approximately 460 volts instead of the 250-volt supply for the transformer-coupled case.

POWER OUTPUT AND PLATE EFFICIENCY. In the operation of vacuum tubes as amplifiers (also as oscillators, or detectors) at low power levels it is usually desirable to get as much *output signal power* as possible from a given tube regardless of the *plate efficiency* (the ratio of the output signal power to the input plate supply power). Under such conditions there is usually one of two parameters limiting the power output, namely, the input signal voltage in amplifiers and detectors, or the plate supply potential in the case of all vacuum-tube operation.

When the input signal voltage is the limiting factor, and the tube is operating over a linear part of its characteristic (for detectors this means linear with respect to its detection characteristic—see Detectors, p. 7-76), the vacuum tube can be considered a generator with an internal resistance. Then the maximum power output occurs when the load is a pure resistance equal to the internal resistance.

Frequently the input voltage available is large enough so that for the given plate supply voltage the introduction of distortion is the limiting factor. Using the Taylor's series development of the plate current up to and including the second power term, Warner and Loughren (*Proc. I.R.E.*, Vol. 13, 709 [1925]) showed that under these conditions the maximum undistorted power output is obtained when the load was resistive and equal to twice the internal resistance. Experiment has generally shown this to be approximately correct for amplifiers using low amplification triodes in class A operation but that no such simple rule applies in high-mu triodes, tetrodes, pentodes, and class B operation of triodes. The optimum value of load resistance is usually determined experimentally and forms part of the operating data furnished by the tube manufacturer.

When vacuum tubes are operated at high power levels it is usually economically necessary to consider the plate efficiency. This is given by the formula

$$\text{Eff.} = \frac{\dfrac{1}{T}\displaystyle\int_0^T i_0 e_0\,dt}{\dfrac{E_b}{T}\displaystyle\int_0^T i\,dt} = \frac{\dfrac{-r_L}{T}\displaystyle\int_0^T i_0^2\,dt}{\dfrac{E_b}{T}\displaystyle\int_0^T i\,dt}$$

where i_0 is the component of the current in the plate circuit having the desired output frequencies and amplitudes; e_0 is the corresponding voltage drop across the load z_L, the resistance component of which is r_L; i is the instantaneous value of the plate current; E_b is the plate supply potential; and T is a complete period of i.

When the current i is given in the form

$$i = I_0 + I_1 \cos \omega t + I_2 \cos 2\omega t + \cdots$$

the plate efficiency becomes

$$\text{Eff.} = \frac{I_1^2 r_L}{2E_b I_0}$$

when the fundamental is the desired frequency of the output power.

PUSHPULL AMPLIFIER. Figure 12 represents a pushpull amplifier, which has several advantages over the single-ended amplifier. A class A pushpull amplifier does not take a pulsating current from the plate supply for the ideal condition, and in practice the variation of supply plate current is negligible. This essentially constant plate current permits the omission of the bypass condenser that is required in a single-ended class A amplifier, shown as C_1 in Fig. 9. Another advantage of the pushpull amplifier is that the d-c magnetizing current in the primary of the plate transformer is balanced out, which simplifies the plate transformer design. (See p. 6-17.)

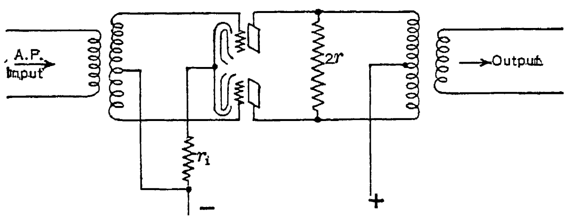


FIG. 12. Pushpull Amplifier

The equivalent circuit of the pushpull class A amplifier is shown in Fig. 13, which is similar to the circuit shown in Fig. 10, except that the resistances of both plates are in series with the load resistance. In order that the load resistance r of 43,000 ohms be obtained for the 56-type tube as discussed for the single-ended amplifier, it is necessary that the primary impedance of the loaded output transformer be 86,000 ohms, which is not very practical for voltage amplifier purposes, because of the large number of turns required in the primary. Another disadvantage of the pushpull voltage amplifier is that one-half of the total input voltage is available for each grid and the increased power output increases the output voltage into a given resistance by only approximately 40 per cent. Therefore, the actual decrease in voltage gain by using two tubes in pushpull-instead of one single-ended tube is about 20 per cent.

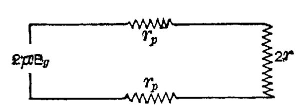


FIG. 13. Equivalent Circuit for Class A Pushpull Amplifier

The calculations of power output of the class A pushpull amplifier may be made by using eq. (10) or (15), if the results are multiplied by 2. The load resistance is calculated by multiplying eq. (12) by 2 so that the transformer primary impedance (plate to plate) is double the value obtained for a single-ended amplifier.

TRIODE VOLTAGE AMPLIFIER. The class A voltage amplifier in general uses a special high-gain type of tube instead of the general-purpose tube such as the 56. The tube characteristic desired is indicated by eqs. (13) and (14). The higher the amplification constant μ, other constants being equal, the higher will be the voltage amplification. Tubes designed primarily for voltage amplifiers of the triode type are the 240, 841, 85, 203A, and others. The general-purpose tube such as the 56 discussed in some detail also has quite extensive application as a voltage amplifier, where interstage coupling transformers are used. In general, the plate resistance of the high-amplification-factor tubes is quite high, so that such tubes are employed principally in resistance-coupled amplifiers. Because of the general use of screen-grid tubes as high-gain amplifiers the triode is becoming less common as a voltage amplifier except where coupling transformers are used.

TRIODE POWER AMPLIFIER. The above discussion applied primarily to the triode as a class A amplifier in which, in general, the grid is not driven into the positive region. With this limitation for grid excitation, it is seen from eq. (10) that the higher the plate-current swing for a given plate-voltage swing the greater will be the power output and correspondingly the greater the plate dissipation. Therefore, a power tube to operate as a class A amplifier, without positive grid swing, should be a low-plate-resistance tube with a correspondingly low amplification constant and capable of relatively high plate power dissipation. The various types of triode tubes primarily designed for the class A power output service are the 31, 45, 2A3, 842, 250, 845, 849, and 848. The type-31 tube is the small battery tube rated as 0.375 watt output; the 848 is a large water-cooled tube rated at approximately 1900 watts output as a class A amplifier.

A typical low-plate-resistance tube designed as a class A power output tube is the 2A3, the plate characteristics of which are shown in Fig. 14. Calculations similar to those made on the 56 tube indicate that about 3.5 watts output can be obtained from the 2A3 with a load resistance of 2500 ohms, a bias of 43.5 volts, a plate voltage of 250, and a plate dissipation of 15 watts at zero output. The plate dissipation decreases from full value at no signal to the full value less the output power with signal. The above 15-watt plate dissipation for the 2A3 decreases to 11.5 watts at full output.

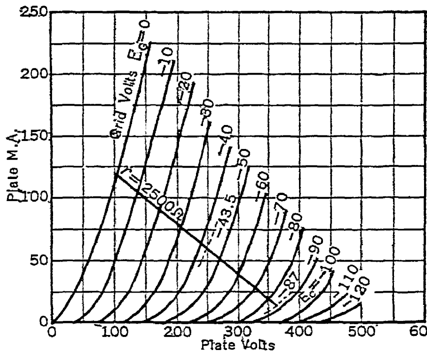


FIG. 14. Type 2A3 Plate Characteristics

PENTODE VOLTAGE AMPLIFIER. The class A pentode amplifier may be used as a voltage amplifier for audio or radio frequencies. The fundamental calculations of power output and determination of load (r) lines is essentially as discussed for the triode amplifier, but certain details of the calculations are different. One of the principal differences is that the amplification factor is much higher than the amplification obtained in practice because the plate resistance, r_p, of the tube is much higher than it is feasible to equal with a load resistance, r. Because of the very high amplification factor of pentodes as well as tetrodes with a correspondingly high plate resistance, eqs. (10), (12), and (13) are much more useful than eq. (14). The transconductance, g_m, for the pentodes and tetrodes is also quite useful in performance calculations.

Referring to the definition and means by which the transconductance was obtained (as given in Section 4), it will be noted that the condition for obtaining data for g_m calculations is that the resistance in series with the plate is zero or at least sufficiently low not to alter the observed plate-current change for a given grid-voltage change. It may be seen from the plate current characteristic of a type-57 pentode, as shown in Fig. 15, that the slope of a load resistance line drawn through the indicated operating point does not alter the plate-current change appreciably for a given grid voltage swing. This tube is used extensively for radio-frequency and audio-frequency high-gain amplifiers. Except where maximum power or voltage swing is desired, it is not necessary to use the plate-current curves to make performance calculations (except to prevent overloading), as such calculations can be readily made from the constants of the tube. For the conditions as given in Fig. 15, the transconductance at −3 volts bias is 1225 micromhos, the plate resistance is given as greater than 1.5 megohms, and the amplification factor is greater than 1500. It is obvious that, with the more or less indefinite amplification factor and plate resistance, the equations involving these constants are of little value in calculating the voltage gain in a 57 amplifier. However, since the g_m of the tube is definite within limits and the above constants are high, the plate-current change for a given signal E_g is: $E_g g_m = I_p$ and $I_p r = E_p$. Therefore, the voltage gain is

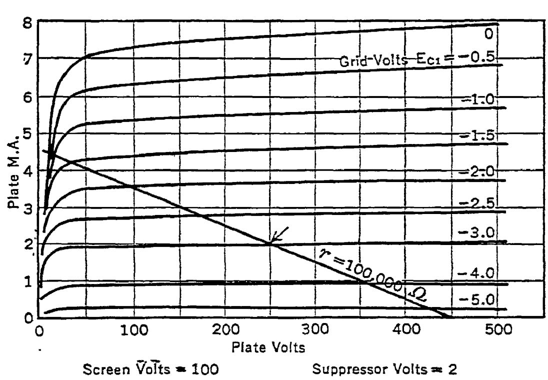


FIG. 15. Type 57, Plate Characteristics

$$\frac{E_p}{E_g} = g_m r = \text{Voltage gain} \tag{16}$$

For the assumed conditions and a load of 100,000 ohms the voltage gain is

$$1225 \times 10^{-6} \times 100{,}000 = 122.5$$

The above value of voltage amplification is approximately 6 per cent high if the plate resistance of the tube is 1.5 megohms, but such corrections are usually unnecessary unless great accuracy is required, and then it is advisable to measure the transconductance of the tube, as well as the plate resistance, at the point of actual operation. A decrease of load resistance to 50,000 ohms reduces the gain to approximately 50 per cent of the above calculated value, and the accuracy is greater because of the lower value of load resistance relative to the actual plate resistance.

In an audio amplifier it is possible to attenuate the high audio frequencies considerably if the load or coupling resistance is sufficiently high so that the shunting capacitances appreciably lower the impedance of r. These capacitances are the plate to suppressor grid capacitance of the 57 amplifier plus the capacitance of the tube or device to which the amplifier is coupled. The output capacitance of a 57-type tube as an audio amplifier is about 6.5 $\mu\mu f$, and the input capacitance to a similar following stage of amplification is about 5 $\mu\mu f$ or a total of about 12 $\mu\mu f$, the impedance of which is about 1.3 megohms at 10,000 cycles, which would not appreciably affect the above 57 amplifier with an r of 100,000 ohms, at the higher audio frequencies. As commonly used the voltage output from a 57 amplifier is so small that little non-linear distortion is introduced.

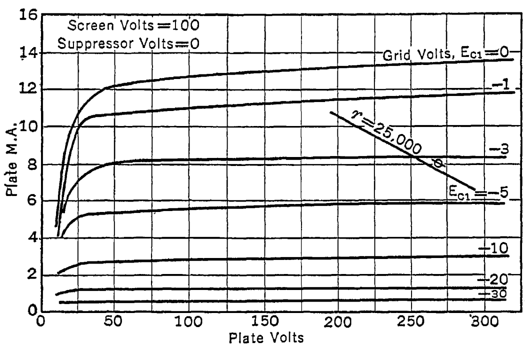


FIG. 16. Type 58, Plate Characteristics

If the tube is used as a radio-frequency amplifier, the plate impedance is obtained by a parallel-resonant, or an equivalent, circuit in series with the plate, and for known equivalent series impedance of the plate circuit the voltage gain may be calculated as indicated above. Since the tube resistance is very high, the selectivity of the amplifier for tuned output is essentially the selectivity of the tuned plate circuit.

VARIABLE-GAIN PENTODE VOLTAGE AMPLIFIER. The 57 tube just discussed is not well adapted to use as a variable-gain control tube by changes in the bias. A tube designed primarily for a bias-controlled variable-gain amplifier is known as the type-58 tube, the characteristics of which are shown in Figs. 16 and 17. Since the plate resistance of the tube is approximately 800,000 ohms or more, the gain as calculated by eq. (16) is quite accurate and the gain decreases as the negative grid-bias increases because the transconductance decreases for increased bias. The distortion in the output of this tube is greater than for the 57-type tube, and for this reason the 58-type is not generally used for an audio voltage amplifier. If the plate circuit of the 58 is tuned and the grid-voltage swing is not too great, little distortion results from the use of this tube in the radio-frequency system of a receiver. The bias may be supplied from an automatic volume-control system, the output of which supplies a bias that is proportional to the carrier value of the received signal, or may be controlled manually, or by a combination of the two. The type-58 or other tubes with the approximate exponential grid voltage vs. plate current characteristics (such as types 34, 35, 39, 78, 6C6) are almost universally employed in receivers so that a simple bias control may be used to control the sensitivity of the radio-frequency amplifier. (See Ballantine and Snow, *Proc. I.R.E.*, Vol. 18, No. 12, p. 2102 [December, 1930].) Such a system permits an automatic volume control to determine the output level to the audio system so that all stations above a predetermined level and with a given percentage of modulation will be reproduced at essentially the same volume.

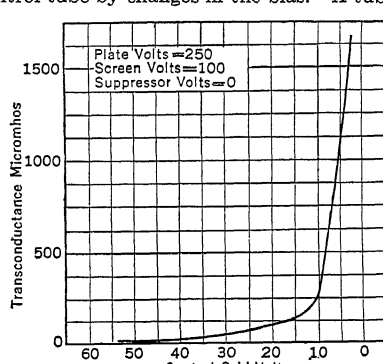


FIG. 17. Type 58, Transconductance

It will be noted that, of the tubes listed, all are pentodes except the early exponential tube, type 35. The principal reason for the pentode construction is that the plate voltage swing is not limited by the screen-grid voltage for the pentode as for the screen-grid tube.

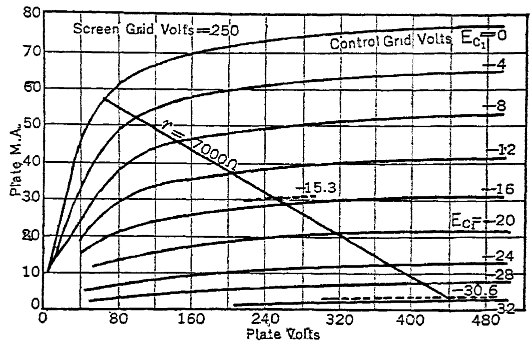


FIG. 18. Type 47, Plate Characteristics

PENTODE POWER AMPLIFIER. This latter characteristic of pentode tubes permits comparatively large power outputs from tubes designed primarily as audio power output tubes. The power pentode tubes are used principally as audio power amplifiers in radio receivers; however, the power-type pentode designed for use in transmitters is gaining favor, but not for class A amplifiers.

The first power pentode tube designed for the output system of receivers was the type 47, but it is being replaced by the indirectly heated cathode pentodes such as types 2A5, 42, and 41. The characteristics of the 47, however, are typical of the power pentodes. The plate characteristics for the 47 are shown in Fig. 18, and power-output calculations for full grid swing may be made by using eq. (10). The load line drawn through the operating bias of 15.3 volts at 250 volts on the plate and screen grid represents 7000 ohms and is approximately the load resistance for maximum power output for full grid swing and minimum distortion. A typical circuit for the pentode class A power amplifier is shown in Fig. 19, for the 47-type tube. The calculated power output according to eq. (10) is approximately 2.4 watts. Two tubes may be used in push-pull similar to Fig. 12. The desirable characteristic of the pentode tube is its power sensitivity, that is, the ratio of power output per volt of signal applied to the grid. This ratio is 3 to 4 times as large as for triodes, but the pentode power tube has several peculiar characteristics which are not particularly desirable.

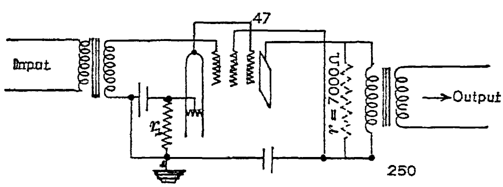


FIG. 19. Circuit for Single Type-47 Audio Amplifier

Distortion in Pentode Audio Power Amplifier. Referring to Fig. 18, it will be noted that the plate-current increase for a peak grid swing of 15.3 volts (the value of the bias) is practically equal to the plate-current decrease for the same grid-voltage swing in the negative direction along the 7000-ohm load line. Therefore, the second harmonic is essentially zero, but it can also be seen that the plate-current change per volt change in bias near the extreme swings of the bias decreases, which produces a flat-topped wave. The predominating harmonic in such a wave is the third, but other harmonics may be present, their value depending on the flatness of the top of the output wave. (See p. 5-47.) It is usually much simpler and more accurate to measure the harmonic content under actual operating conditions than it is to calculate it. All the unknown factors, such as inefficiency of output transformer and lack of proper plate supply filtering, are accounted for in the measured values; their effects are very difficult to determine and express mathematically. Experimental curves of percentage distortion of a 47 tube are shown in Fig. 20.

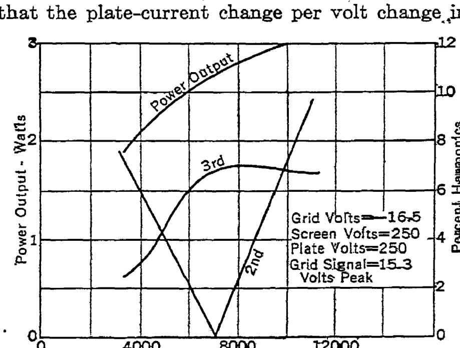


FIG. 20. Type 47, Output Characteristics

BIDIRECTIONAL AMPLIFIERS—REPEATERS. All the circuits considered above amplify in one direction only and will not pass appreciable energy in the other direction.

There are some applications, notably two-way telephony over wires, where it is desirable to amplify in either direction. (See also Section 17, p. 39.)

Figure 21 shows a schematic diagram of a circuit which amplifies signals coming from either end of the circuit, line West or East, and distributes the amplified signals equally between the two lines. Inclosed within the dotted line is a three-winding transformer; as shown in article 7, p. 6-12, if the proper circuit adjustments are made, energy coming from either line is divided by the three-winding transformer equally between the input circuit (connected across A–B) and the output circuit (inductively connected). The half entering the output circuit is dissipated as heat; but the half entering the input circuit is introduced into the grid circuit of a vacuum tube through a step-up transformer and a potentiometer. This is amplified by the tube, or tubes, and appears in the output winding of the three-winding transformer where it is again divided equally between East and West lines. Thus, although the incoming energy is twice divided, so that only one-fourth of it is utilized, the power amplification of the tube, or tubes, may easily be several hundred, so that the net result is a great increase in energy over the original.

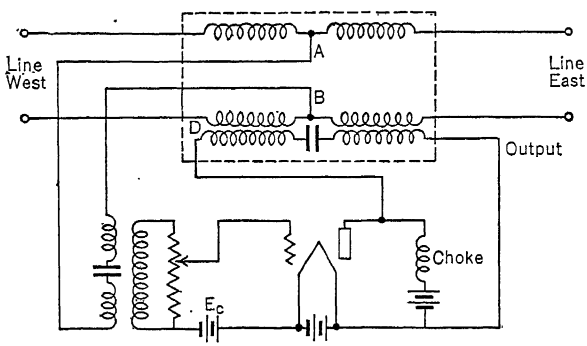


FIG. 21. Bidirectional Amplifier

The chief practical difficulty with this circuit is that the impedance of lines West and East must be identical at all frequencies, or energy will be fed by the three-winding transformer from the output to the input circuit and the tube will furnish sustained oscillations, or "sing."

To avoid the requirement of working between similar impedances a circuit such as is shown in Fig. 22 is used. Here a network of resistors, inductors, and condensers is designed to balance the external impedance on each side of the amplifier. The principle of operation is the same as above, the left-hand tube amplifying signals entering from the line West and dividing the energy between line East and the accompanying network (N).

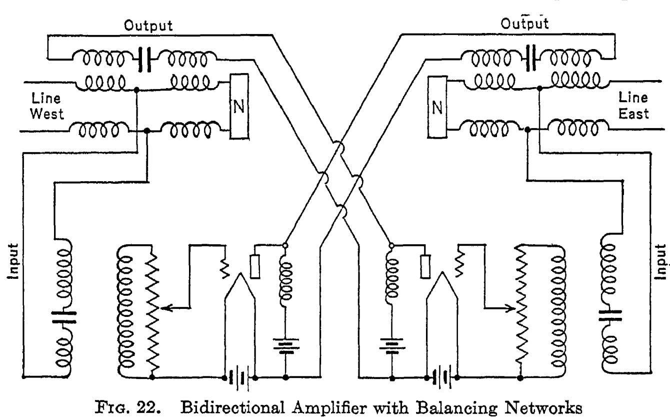


FIG. 22. Bidirectional Amplifier with Balancing Networks

Similarly, signals entering from line East are amplified by the right-hand tube and divided between line West and its network. An impedance unbalance between either line and its associated network causes some of the energy of the output circuit to be fed by the three-winding transformer into its input circuit, where it is amplified and fed into the other transformer. An unbalance here will similarly cause some of the energy to find its way back through the original path. Thus it is seen that if both lines are poorly matched by their networks the amplifier may act as an oscillator, but if either line is *perfectly* balanced by its network no sustained oscillations will be produced.

Amplifier circuits of this type are in extensive use by telephone and telegraph companies on their long lines. They are called *repeaters*.

SUMMARY FOR CLASS A AMPLIFIERS. The foregoing discussion of the class A amplifiers and the various types of tubes for particular services assumed that the grid swing or excursions were always in the negative region, and that the amplifier behaved in general according to the definition of the class A amplifier, as given at the beginning of this section on amplifiers. There are a few facts about the characteristics of class A amplifiers which should be kept in mind when designing or studying such devices; some of them are listed below.

1. The maximum voltage output of a class A amplifier measured at the plate of the tube is limited by the plate-voltage limitations of the tube and the plate load, r.
2. The maximum power output from a class A amplifier is limited by the plate dissipation at zero output and the minimum instantaneous plate voltage for peak positive grid swing to zero, at which the plate current is approximately double the steady plate-current value.
3. The load impedance into which the tube must work for maximum power output is not a simple function of the plate resistance of the tube but depends upon the two conditions above in cases where the grid swing is not limited. Power-output tubes have high bias and low plate resistance so that grid swings to zero voltage are not the controlling limitation.
4. The power output for a comparatively low and limited grid swing, that is, when limitations of 1 and 2 are not effective, is maximum when the load resistance is equal to the plate resistance of the tube. For this same grid-swing limitation, the maximum output voltage or gain approaches the amplification factor of the tube as the load resistance, r, increases toward the value of the tube plate resistance, r_p.

2. CLASS B AMPLIFIERS

GENERAL. Because of the relatively low power output of class A amplifiers and the fact that the plate dissipation is maximum for no-signal conditions, the overall efficiency of such an amplifier is very low. In applications where considerable power is desired either at radio or audio frequencies, the size of the tubes and the cost of plate power supply, per watt output for class A amplifiers, increases so rapidly that such amplifiers are not used. In the broadcast receiver it is not economical to obtain power outputs greater than about 5 or 6 watts without resorting to class B amplifiers or some combination of the class A and B amplifiers.

The grid swing in the class A amplifier is usually limited to the negative region for the entire input cycle because, in general, grid swings into the positive region result in plate-current distortion, chiefly because of the large external grid circuit impedance and limitations in plate-current swing.

According to the definition of a class B amplifier, the bias is such that the operating plate current is small, so that for the no-signal condition the plate dissipation is low. Therefore, the grid swing in general is limited only by a non-linearity of plate current and grid voltage when the grid swings positive from the operating point. Almost any three-element tube and some pentodes may be used in a class B amplifier. However, some tubes are to be preferred for reasons that will be evident after the various calculations are made from the various sample tube characteristics as shown and discussed below.

LOW-POWER AUDIO AMPLIFIERS. Since self-bias, as obtained in a pushpull circuit as shown in Fig. 12, depends upon the signal (because the *average* plate current in a class B amplifier depends upon the signal) and since there is no other convenient means for supplying fixed bias in the usual receiver, low current at zero bias is very desirable in tubes for use in class B amplifiers. The plate characteristic curves for such a tube are shown in Fig. 23, with the corresponding grid currents for various grid voltages plotted against plate voltage. Experience has indicated that a plate load resistance of about 2000 ohms for this tube with 400 volts on the plate is approximately the optimum value for power output and plate-circuit efficiency. The 2000-ohm load line is drawn through the operating point to obtain data for power-output calculations and grid-current values to determine the input resistance of the grids. Data can be obtained from this set of curves to replot curves of plate current and grid current against grid voltage, which are more useful for performance calculations than the curves in Fig. 23. Such curves are commonly called dynamic transfer characteristics.

The data for these dynamic transfer characteristics may be optionally taken directly from meter readings in a circuit such as shown in Fig. 24. Here are plotted various curves for different load resistances. Except at points near the zero bias axes, that is, for a grid swing of about 5 volts, these curves represent the dynamic performance of the 46 tube for various load resistances for one half-cycle. Another, similar tube is connected in pushpull

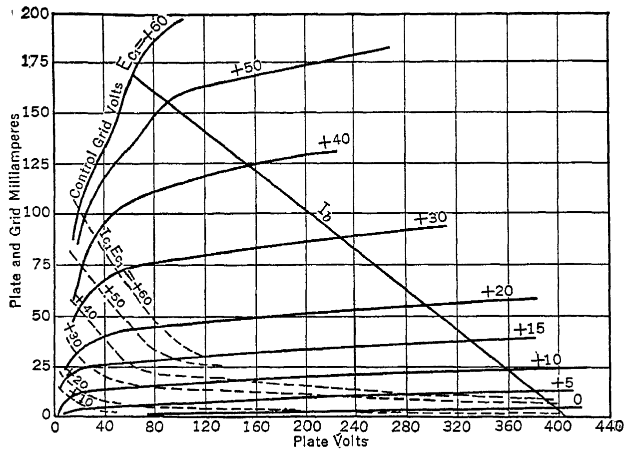


FIG. 23. Type 46, Plate Characteristics

in order to supply power for the other half-cycle. If the two tubes have similar characteristics and the pushpull input and output transformers are well balanced, the output wave will be symmetrical and only odd harmonics will be present in the output for a sinusoidal input wave.

The power-output calculations for the 46 from the curves in Fig. 24 are made like the power-output calculation for the class A amplifier, except that the plate current $I_{\max}$ (the peak plate current read from the curves) is the total change in plate current from the operating point of essentially zero plate current. Since $I_{\max}$ is the peak value of the output wave, and if little distortion is present for sinusoidal inputs, the power output for the two tubes is $I^2_{\max}r/2$. The average plate-current input for two tubes at a constant plate supply voltage, E_b, for approximately full power output is $I_{\max}2/\pi = 0.637I_{\max}$. Since the plate voltage is constant the power input to the plates of the two tubes is $0.637I_{\max}E_b$. Therefore, the plate efficiency is

$$\text{Pl. eff.} = \frac{I_{\max}r}{1.27E_b} \tag{17}$$

and the plate dissipation per tube is

$$\text{Pl. loss} = \frac{0.637I_{\max}E_b - 0.5I^2_{\max}r}{2} \tag{17a}$$

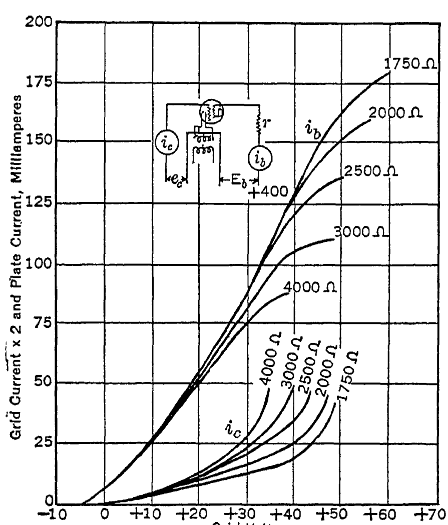


FIG. 24. Type 46, Dynamic Transfer Characteristics

The maximum theoretical efficiency is obtained when the peak a-c plate voltage is equal to the d-c plate voltage and, according to eq. (17), is 78.5 per cent.

The above equations are quite accurate at full or nearly full grid swings; the accuracy decreases with lower grid swings, the decrease in accuracy being a function of the no-signal or standby plate current. For the theoretical condition for a class B amplifier of zero plate current at no signal, they are accurate for all grid swings.

If the maximum value of plate current is taken from the curves in Fig. 24, for the 2000-ohm load, sample calculations are as follows:

$$\frac{0.150^2 \times 2000}{2} = 22.5 \text{ watts output}$$

$$\frac{0.150 \times 2000}{1.27 \times 400} = 59 \text{ per cent plate efficiency}$$

$$\frac{0.637 \times 0.150 \times 400 - \dfrac{0.150^2 \times 2000}{2}}{2} = 7.9 \text{ watts plate loss per tube}$$

It should be noted that the load resistance used is the resistance the tube works into during the half-cycle it functions so that the plate to plate equivalent load on the primary of the output transformer is $4r$.

INPUT RESISTANCE. The non-linear grid current in class B amplifiers for audio frequencies is a principal difficulty to overcome. The plate-current grid-voltage curves, as shown in Fig. 24, are for perfect regulation of grid voltage or the equivalent of zero resistance in series with the grids. An essentially zero input resistance to the grids, however, is not practical to attain, but values can be obtained which are quite low compared to the minimum instantaneous resistance of the grids of the class B amplifier tubes.

The slope of the grid-current curves in Fig. 24 at any particular grid voltage gives the instantaneous input resistance of the amplifier. The minimum value of the grid resistance occurs at a maximum grid swing of about 45 volts positive. The slope of the grid-current curve at this point represents approximately 500 ohms, while at points below 45 volts on the grid, the grid resistance is about 1600 ohms, and at near the zero axis, the resistance is still higher for each tube. However, the grid current is not zero at zero bias, so that the input resistance of the two tubes over the region at which grid current flows to each tube is the resistance of the two tubes in parallel. Therefore, to drive the grids of two 46-type tubes properly as audio amplifiers, the amplifier stage supplying voltage to the grids of the 46's works into a load resistance of approximately 4500 ohms for each tube, or a combined resistance of approximately 2250 ohms for a short period during which each tube draws grid current. When one tube ceases to draw grid current, because its grid is too negative, the input resistance rises very rapidly to approximately 4500 ohms, after which it decreases gradually to 500 ohms.

Because of this erratic change of input resistance to the 46's as class B audio amplifiers, it is very important to keep the equivalent impedance in series with each grid a minimum. The low-impedance requirement is met by using low-plate-resistance driver tubes, preferably in pushpull with a transformer *step-down* ratio as great as the plate-voltage swing of the driver tubes will permit; also the coupling transformer leakage reactance and resistance must not appreciably affect the impedance in series with the grids.

The input requirements may best be understood by referring to the circuit for a class B audio amplifier shown in Fig. 25, with the various equivalent circuit constants inserted.

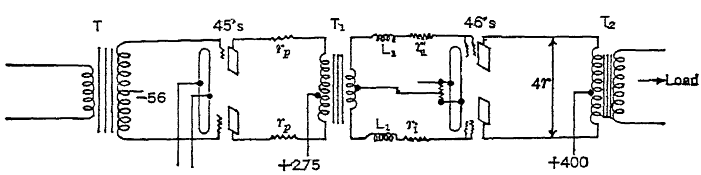


FIG. 25. Circuit and Equivalents of a Class B Audio Amplifier

The combination of tubes shown is perhaps the most practical combination for power outputs of approximately 25 watts. The transformer T is merely an interstage voltage transformer to supply audio voltage from the source to the grids of the type 45 tubes as class A amplifiers. For the grid and plate voltages as shown, the peak plate-voltage swing is approximately 165 volts each side of the operating point for load resistances of the order of 10,000 ohms, which would be a peak voltage of 330 volts on the primary of transformer T. Since according to Fig. 24 the grid swing for the 46 tube, for full output of about 25 watts, is approximately 55 volts, the ratio of the transformer T_1 is the ratio of 330 to 55 or 6 to 1, as measured from plate to plate of the driver tubes to each grid of the 46 tubes. This transformer has an impedance ratio of 36 to 1. The plate resistance, r_p, of the 45 tube is approximately 1800 ohms per tube so that the equivalent resistance in series with the primary of T_1 is $2r_p$, or 3600 ohms for the above case. Therefore, the equivalent resistance in series with the grid of each 46 tube is (in addition to transformer losses)

$$\frac{2r_p}{(\text{Turn ratio})^2} = r_1 = 100 \text{ ohms} \tag{18}$$

The actual impedance is usually about 10 per cent higher than the above calculated value because of transformer losses, but such losses are usually allowed for by assuming a grid swing about 10 per cent or more higher than the actual swing needed for a given power output. In addition to the equivalent resistance, r_1, in series with the grids of the 46 tubes, there is an equivalent inductance, L_1, which is the equivalent leakage reactance in series with one side of the secondary of transformer T_1. As explained above, the change in input resistance of the 46 tube is very rapid over certain parts of the audio cycle which results in current flow to the grids at harmonic frequencies much higher than the fundamental frequency. The frequency at which these grid currents flow may be 5 to 10 times the frequency of the fundamental, depending upon the frequency and amplitude of the signal. With currents flowing at such frequencies to the grids of the 46's, it is evident that the equivalent inductance in series with the grids due to leakage reactance of the transformer T_1 must be low in order that the grid voltage at these frequencies will be low. A high leakage reactance in the input or driver transformer manifests itself in the form of a fuzzy or a ragged output wave over certain portions of the cycle, and it may cause very high transient voltages in the plate circuit, which are responsible for many output transformer breakdowns and the tube failures.

OUTPUT CIRCUIT REQUIREMENTS. Assuming a low-impedance circuit to the grids of the 46's, which results in very little distortion of voltage supplied to the grids, there is still a source of distortion in the plate circuit because of the deviation of the plate-current curves from a straight line. This deviation is shown in Fig. 24, and if the points about the zero axis are plotted on a much larger scale and the slopes of the two plate currents added over the period during which plate current flows to each tube, it will be found that the per cent deviation is perhaps greater at zero than at any other point. In the 838 tube which was specially designed for class B operation there is appreciable plate current at no-signal condition; however, the shape of the plate-current curve is such that when two tubes are operated in pushpull the resultant output current is practically linear. This characteristic of the 46 may be seen by referring to the results of distortion measurements (Fig. 26) as made on an experimental amplifier with constants approximately as shown in Fig. 25. The non-linearity near the zero signal is distinctly shown by the rapid rise of the third harmonic at about 1-watt output, after which one tube ceases to work for most of a half-cycle, resulting in a partial balancing out of the third harmonic for certain power outputs.

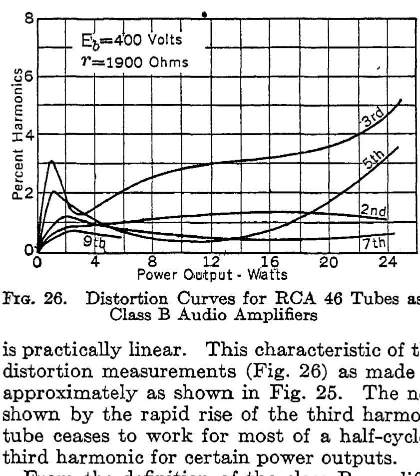


FIG. 26. Distortion Curves for RCA 46 Tubes as Class B Audio Amplifiers

From the definition of the class B amplifier, and from eq. (17), it is evident that the plate current to the class B audio amplifier tubes is approximately proportional to the grid excitation and that the efficiency increases to a maximum at full output. Referring to Fig. 24, it is seen that, if the plate voltage decreases because of poor regulation of the plate supply, the peak plate currents will not be reached as indicated by the curves. Therefore, the plate supply voltage should have good regulation and should be by-passed well for instantaneous peak plate currents. (See p. 7-108.)

The shape of the plate current curves in Fig. 24 cannot be used to calculate distortion by the simple formulas given above for the second and third harmonics. The fact that the output resistance from the driver stage is about 20 per cent of the grid resistance at near the peak swing alters the plate-current curve appreciably at the upper limits of plate current, which complicates the mathematical determination of harmonic outputs. Because of the complicated nature of the procedure in calculating harmonics, it is much easier and more accurate (for small amplifiers) to analyze the output wave, from an experimental amplifier including the driver, for harmonics by means of a reliable voltage analyzer.

LIGHT-WEIGHT CLASS B AUDIO AMPLIFIERS. Because of the high efficiency of the class B audio amplifier its use is very desirable in a battery receiver, in an automobile receiver, or where space and weight are limited. The use of small tubes for a given power output is represented by a miniature 6AU6 driving a single 1635 class B output tube. The output power from this combination is of the order of 12 to 15 watts audio power with low distortion and with very low plate drain by both the driver and output tubes.

As explained above, bias for a class B amplifier is difficult to obtain and keep constant because of the summation of grid-current and plate-current peaks that must flow through a self-bias resistor and because of the grid-current peaks if a separate bias is used. Therefore the 1635 tube was designed for zero bias at a maximum of 400 volts plate supply.

A 4000-ohm load line may be drawn through a point at zero bias and 400 volts supply to the plate on a family of published plate characteristics for the 1635. The dynamic transfer characteristics are shown in Fig. 27 for a 4000-ohm load resistance. The grid

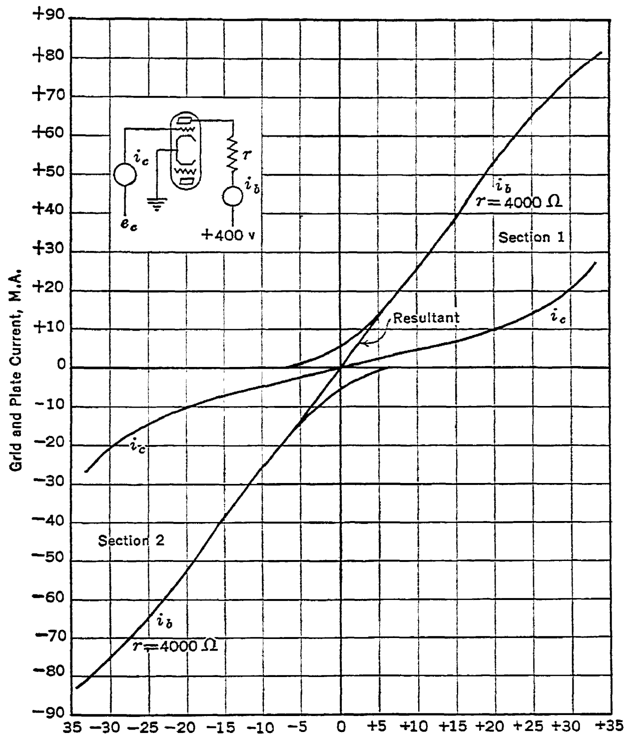


FIG. 27. Dynamic Transfer Curves for a 1635 Tube as a Class B Audio Amplifier

current vs. grid voltage is also shown in Fig. 27. A careful examination of the plate-current curve in Fig. 27 will show that the current is quite low at zero bias but trails out before cutoff similar to a variable-mu tube.

At zero bias about 6 ma plate current flows to each plate so that each plate contributes to the output. Since the plate resistance of the 1635 is high the output power from each plate is measured by plate-current deviation from its zero value through the load resistance. In pushpull operation if the plate current of one tube is increasing the other is decreasing, and since the phase of the currents is reversed in the output transformer the output current is given by the difference in plate current to the two plates. The true output-current curve is then obtained and plotted in Fig. 27 as the resultant output current. This resultant-output-current curve is more nearly linear for the 1635 tube than for other double tubes because of the special grid construction. Because of the special grids the grid current at zero bias is lower so that the sum of the grid resistance of the two tubes is higher than for other double plate tubes during the time grid current is flowing to both grids. This higher input resistance to the tube for small signal swings reduces so-called high frequency or "fuzz" type of distortion.

At a representative peak plate current of 80 ma in Fig. 27 the output is $I^2R_P/2 = 12.8$ watts. The grid peak swing is +32 volts for the above peak plate current, and the grid current is 25 ma peak. These peak grid values of current and voltage permit the driver calculations. The slope of the grid-current vs. grid-voltage curve at +30 to +32 volts represents a resistance of 500 ohms, increases to about 1500 ohms at about +6 volts on the grid, and decreases again to about 700 to 800 ohms near the zero axis because of grid current flowing to both grids. The above variation of input resistance necessitates a "stiff" driver.

The plate resistance of the 6AU6 pentode to be used as a driver is quite high so that degeneration is used to lower its output resistance. The circuit for the 6AU6 driver and the 1635 output amplifier is shown in Fig. 28.

From the published plate characteristics of the 6AU6 under the voltage conditions shown in Fig. 28 it will be seen that the bias for class A operation is about 1.5 volts at a plate current of about 7 ma. At double plate current and a signal swing to zero bias the plate swing is about 225 volts. A peak swing of about 32 volts is needed on the grids of the 1635 so that the step-down ratio of the driver transformer to each grid is about 7 to 1.

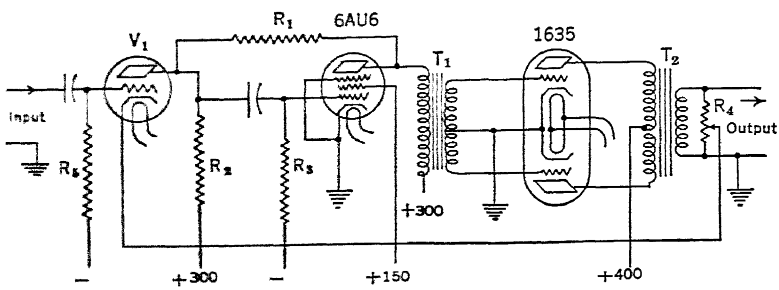


FIG. 28. Driver and Output Circuit for a 1635 Class B Audio Amplifier

The maximum current output from each side of T_1 is then 7 ma × 7 = 49 ma, which is more than ample driving capability for a needed peak grid current of 25 ma. As noted above, the variation of grid resistance or load on T_1 was 500 ohms to about 1500 ohms. The input resistance should be as low as 10 per cent of the minimum grid resistance to keep the input distortion to a low value.

The g_m of the 6AU6 is about 5000 micromhos, and, if it is assumed that 1 volt peak were applied to the plate of the 6AU6 and that 0.2 of a volt peak gets back to the grid of the 6AU6 through R_1, the effective output impedance of the 6AU6 to one side of the secondary of T_1 is:

$$\frac{1 \text{ volt}}{0.2 \times g_m} \times \frac{1}{(\text{Ratio of } T_1)^2} = \frac{1000}{49} = 20.4 \text{ ohms}$$

The effective internal resistance of the transformer, T_1, to each side of its secondary may be about 20 ohms for a good design so that the total input resistance in series with 1635 grids is about 40 ohms. This low value of resistance will effectively reduce input distortion to a low value.

Referring to Fig. 28 if the combined plate resistance of V_1, R_2, and R_3 is 50,000 ohms, R_1 in the above calculations for feedback voltage will be 200,000 ohms. Therefore, 20 per cent of the peak plate swing of 225 volts on the plate of the 6AU6 will require about 1 ma of the 6AU6 plate-current peak swing and apply 45 volts audio on the plate of the tube V_1. In this case V_1 must supply an equivalent voltage of 45 volts at its plate to balance out the feedback voltage plus a peak voltage of about 2 volts to drive the grid of the 6AU6.

For a 1635 plate load of 4000 ohms per plate the plate to plate load for T_2 is 16,000 ohms. The transformer T_1 must have low leakage reactance to prevent fuzz-type distortion as explained for T_1 in Fig. 25. A check of constants for the 6AU6 as a degenerative feedback driver for the 46's in Fig. 25 indicates that the 6AU6 is essentially as good a driver as or better than the two 45's. This indicates the effectiveness of the addition of a feedback resistor R_1 in Fig. 28.

Some degeneration may be fed back from a resistance divider, R_4, across the output of T_2 to the cathode of V_1 as shown in Fig. 28 to decrease the distortion of the system further. The distortion vs. output power of the circuit shown in Fig. 28 is shown in Fig. 29. The feedback for the overall distortion curve was an amount required to reduce the overall gain about 2 or 3 to 1.

Other sharp cutoff pentodes such as the 6SJ7 may be used as drivers. Battery tubes of the 1.4 volt series may be used in combinations to obtain high outputs for given input powers.

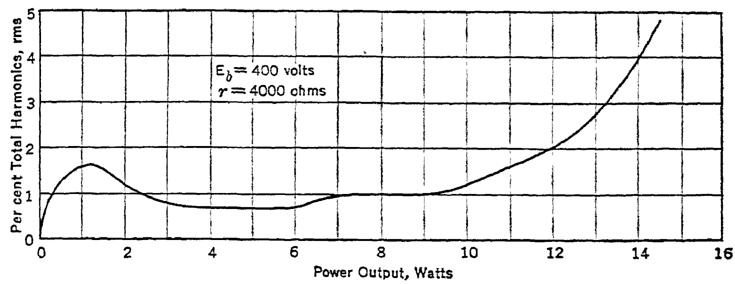


FIG. 29. Overall Distortion of the Circuit of Fig. 28 Using the 1635 as the Output Tube

MEDIUM-POWER AUDIO AMPLIFIERS. A set of curves similar to the curves in Fig. 24 is shown for the UV849 tube in Fig. 30. It will be noted from these curves that the class B operating bias is approximately -140 volts and that the instantaneous slope of the grid-current curves is positive and negative with magnitudes as low as 500 ohms. Considering the relatively high grid voltages at which these low instantaneous grid resistances occur, these tubes are hard to drive without appreciable distortion. However, two UV845-type tubes operating at full rated voltages may be used to drive these tubes successfully with an equivalent resistance of approximately 100 ohms in series with the grids of the UV849 tubes.

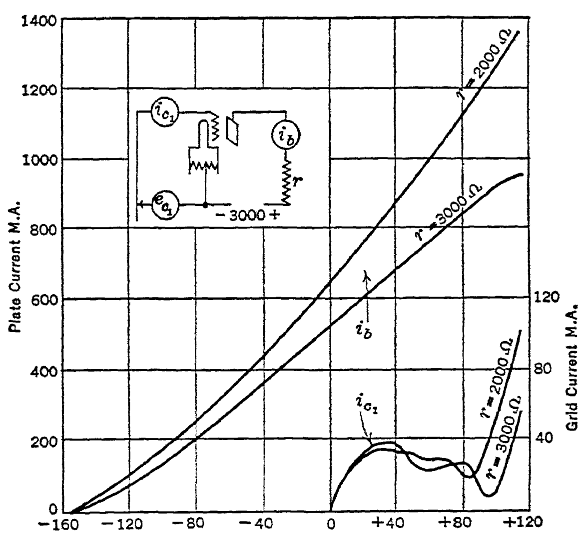


FIG. 30. Dynamic Transfer Characteristics of 849 Tube

By applying eq. (17) to calculate the various powers from data obtained from Fig. 30, it is found that the maximum power output for the 2500-ohm load line is about 1350 watts for two tubes, the plate loss is about 350 watts per tube, with a plate-circuit efficiency of approximately 65 per cent. These calculations do not take into account the transformer losses. Such audio power outputs are used principally to plate-modulate a class C radio-frequency output amplifier.

HIGH-POWER AUDIO AMPLIFIER. The 500-kw broadcasting transmitter operated at WLW in Cincinnati, Ohio, used high level modulation with a class B audio amplifier capable of approximately 325-kw audio output power. At the time this station was first operated, early in 1934, it was the world's largest broadcasting station. The class B amplifier for the above station used eight type-862, 100-kw tubes, with 12,000 volts on the plates.

As stated above, tubes other than the ones discussed may be used as class B audio amplifiers, but their particular grid characteristics must be known in order that a driver system may be properly designed. The load resistance and plate-current characteristics are also important items to consider in the design of a satisfactory class B audio amplifier.

RADIO-FREQUENCY AMPLIFIERS. The class B radio amplifier functions much the same as the class B audio amplifier as far as the grid currents and plate currents are concerned, except that the plate-current distortion may be greater with a correspondingly higher output power without serious signal distortion. This is because the second and higher harmonics are multiples of the radio frequency and are outside of the frequency range of interest. The principal difference between the two amplifiers is in the input or driver circuits and the output circuits. If the input or driver circuit is tuned, the harmonics are by-passed and do not appear on the grids, which results in a well-regulated or low-impedance driver source, as far as the fundamental frequency is concerned. As in the class B audio amplifier, the regulation of the voltage to the grids of the class B tubes depends upon the ability of the driver tube to supply the instantaneous grid currents directly.

Another factor which reduces distortion to a negligible value is the fact that the plate circuit is timed; thus it will absorb the fundamental component of the plate current and by-pass the harmonics. The power output of the class B radio amplifier may be calculated like the power output for the class B audio amplifier if somewhat greater peak currents are assumed for peak outputs. It would seem from the above that the class B radio amplifier, also known as the linear amplifier, is inherently a simple type of amplifier to design; this is true if it is desired only to obtain a power amplification of the fundamental frequency. However, it is not economical to use such an amplifier as merely a tuned amplifier for voltage or power amplification, because in such cases the class A amplifier would probably be better for voltage amplification and the class C amplifier for power amplification. The real use for the class B radio amplifier is in circuits where a low-powered modulated signal is to be amplified to a higher power level. Referring to the definition of the class B amplifier, one characteristic of the amplifier is that the output voltage is proportional to the input voltage; therefore, if a modulated radio-frequency signal is used to excite the grids of a class B radio amplifier, the output for an ideal case would be a modulated signal identical to the modulated input signal, except that it would be at a much higher power level. The increase in power level is 20 to 100 times, depending upon the degree to which the amplifier is driven.

A typical circuit for such an amplifier is shown in Fig. 31, in which the tuned input circuit $L_1C_1C_2C_3$ is so arranged that C_2 equals C_3 and each has a value with respect to

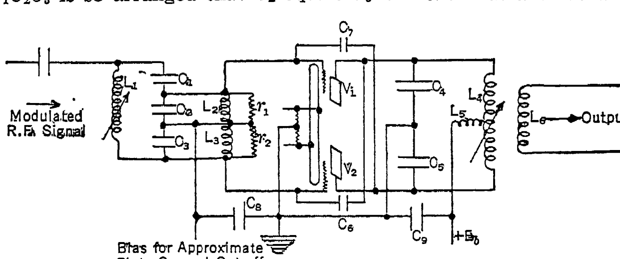


FIG. 31. Circuit for Class B Radio Amplifier

C_1 that permits the desired grid swing of the tubes, V_1 and V_2, for a given modulated signal applied to the input circuit. Inductances L_2 and L_3 are radio-frequency chokes to provide a low d-c resistance path for the grids, and r_1 and r_2 are resistors from the grids to ground to suppress undesirable oscillation. C_6 and C_7 are neutralizing condensers, and the output tuned circuit is $L_4C_4C_5$, in which C_4 and C_5 are equal. The radio-frequency choke L_5 provides a floating center tap for L_4 because the tuned circuit is center-tapped in the condenser branch. The circuit as shown is for a pushpull class B radio-frequency amplifier, but, as stated above, a single tube may be used in a circuit similar to the one discussed, by leaving out the unnecessary parts or sections of the circuit. The type of circuit shown, or its equivalent, was in common use in most broadcasting stations of 1000 watts or more. However, the present trend is toward the use of the more efficient high level modulation system. (See p. 7-74.)

Distortion. As discussed above, there is little distortion from a tuned input and tuned output amplifier as far as the fundamental radio frequency is concerned. Since the primary use of the linear amplifier is to amplify modulated signals without distortion, the linearity of the amplifier becomes a factor which determines almost entirely whether the demodulated output is distorted or not. The linearity of such an amplifier is determined by the ratio of output current, or voltage, against rms input voltage. The curves of Fig. 32 are plotted from actual experimental data obtained on two RCA 846 tubes in pushpull as a class B radio or linear amplifier with 7500 volts on plate. For a bias of -150 volts, the output tank current is approximately linear with respect to the grid-excitation voltage from zero to about 4 amp. If the amplifier is used as a radio-frequency output amplifier in which the input signal is 100 per cent modulated, the normal output

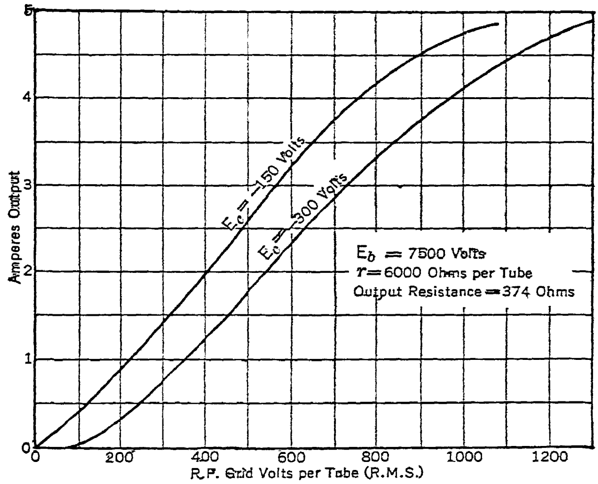


FIG. 32. Characteristics of RCA 846 as Class B R-f Amplifier

tank current would be approximately 2 amp, which corresponds to a grid excitation of 470 volts for approximately 1300 watts output. It will be noted that such high per cent modulation of the excitation voltage in this case causes a deviation from a straight line at both the upper and lower limits of swings so that the predominating harmonic in the demodulated output signal is the third. If a lower bias is used the curve would be more nearly linear from the operating point of 2 amp, but the slope at the upper end would be approximately the same resulting in a second harmonic for 100 per cent modulation.

The flattening of the output tank current near the zero axis, for the curves in Fig. 32, is due to the large bias chosen; however, the flattening of the upper end of the curve may be due to emission or space-charge limitations which limit the peak plate currents. Another factor that often contributes to the flattening of the upper end of the output tank current curve is the inability of the modulated exciter to supply the higher value of average grid current without distortion to the output tubes at the peak values of grid excitation. The driver must have low impedance to the grids of the output tubes to prevent distortion, for practically the same reasons as a low impedance is required for the input to the class B audio amplifiers.

Since the normal carrier value of output current for the class B radio amplifier is approximately one-half the peak value and occurs at one-half the peak value of grid excitation, it can be seen from eq. (17) that the efficiency of the plate circuit is relatively low. (See p. 7-17.) This efficiency is usually 20 to 30 per cent. The carrier power output is only 25 per cent of the peak power output; therefore, except in tubes where peak plate voltage or peak plate power input is limited, four times as many tubes are theoretically required for class B radio amplifiers for a given power output as are required for a class C high level modulated radio-frequency amplifier. Actually the water-cooled tubes are limited by peak plate voltage for the plate modulated class C amplifier so that the number of tubes for a class C amplifier, when these tubes are used, is approximately one-half that required for a class B radio amplifier.

SUMMARY FOR CLASS B AMPLIFIERS. Some of the more important requirements to meet in the design and operation of class B amplifiers and some of their characteristics may be summarized as follows.

1. The input system to the class B audio amplifier must have low impedance, especially if the instantaneous values of grid current change rapidly either negatively or positively. The slope of the grid-current curve or its rate of change per volt applied to the grid is the grid resistance at the point taken. The equivalent inductive reactance in series with each grid because of driver transformer leakage is a part of the impedance in series with the grids; it must be kept very low if the instantaneous values of grid resistance are low and change rapidly over the audio cycle.
2. Failure to meet the low-impedance input requirements usually results in appreciable distortion, and generally the output transformer is subjected to high peak voltages which may damage the transformer or the tubes.
3. The load resistance the tube works into is one-fourth of the calculated plate to plate impedance of the loaded output transformer. This applies to radio-frequency as well as audio-frequency amplifiers.
4. The load line for the class B amplifier for radio and audio frequencies should be as high a resistance as possible consistent with grid current peaks and space-charge limitation in order that the plate efficiency may be maximum. Somewhat higher peak plate currents may be assumed for the class B radio amplifier, because of the tuned plate circuit.
5. The class B audio amplifier is the most efficient type of amplifier for audio frequencies, and may attain a plate efficiency of 65 to 70 per cent at full output power for some of the larger tubes.
6. The class B radio amplifier has poor plate efficiency, which is usually 20 to 30 per cent, so that a relatively high tube power capacity is required for a given output.
7. The class B radio amplifier also requires a well-regulated or low-impedance driving source, but the requirements are not as strict as the driver requirements of the class B audio amplifier, because the radio amplifier may have a tuned input circuit.
8. The plate supply for the class B audio amplifier must have good regulation, especially if tubes with high bias are used. The plate supply regulation is not so important with the class B radio amplifier because the average plate current is essentially constant. Sufficient plate supply by-pass capacitance is necessary in either amplifier to maintain constant plate voltage over the audio cycle.

3. CLASS C AMPLIFIERS

GENERAL. The class C amplifier by definition is an amplifier that is biased considerably beyond plate current cutoff, so that plate current flows for a period less than 180 electrical degrees, during the time the grid swings in a positive direction from its normal value of bias potential. In general, the grid is driven to the point of saturation, that is, to a point at which the output voltage is no longer proportional to input voltage. Under this condition the efficiency of the amplifier is approximately maximum for any given plate-circuit conditions, the output voltage is proportional to the d-c plate voltage, and the output power is proportional to the square of the d-c plate voltage.

Since, in general, appreciable power is required to drive the class C amplifier because of the relatively high positive grid swings, this type of amplifier is essentially a power-amplifier device. Because of the distorted plate current that flows during only a part of a half-cycle it is necessary to tune the output or plate circuit of the amplifier if an undistorted output wave is desired. Therefore, the class C amplifier is essentially a single-frequency amplifier and is used as the output device for radio telegraphy or high level modulated telephony transmitters.

CIRCUIT CALCULATIONS. Since the class C amplifier can be used only as a tuned plate amplifier except where distortion of the output wave is permitted, it is obvious that the application of this amplifier is much more limited than the applications of the class A or class B amplifiers. In the class C amplifier it is difficult to predict the performance of a tube from its static characteristics. As long as the plate circuit load, r, has a sufficiently low value to limit the minimum instantaneous plate voltage to a value such that the plate current is essentially half sine waves, the power output and other characteristics of the circuit and tube may be calculated in much the same manner as for the class B audio amplifier. However, under these conditions, the plate-circuit efficiency is usually less than 65 per cent. When the load resistance is high the minimum instantaneous plate voltage may reach or approach zero, so that essentially no plate current can flow at the instant, because the plate voltage is near zero.

To illustrate better the above features of the class C amplifiers as well as the general circuit features, it is well to discuss the typical class C amplifier circuit such as shown in Fig. 33. The proper constants to use in the circuit depend upon the tube used, general class of service, and the values of maximum grid and plate currents allowed. For the 204A tube the maximum allowed constants are: d-c plate voltage, 2000 volts; d-c plate current, 275 ma; plate dissipation, 167 watts; d-c grid current, 80 ma.

If a modulation factor of 1 is applied to the plate voltage, it is evident that at peak upward modulation the instantaneous plate voltage will be 4000 volts. In order that plate current cutoff may be obtained at 4000 volts on the plate, so that the tube may continue to operate as a class C amplifier at this voltage, the bias required is approximately −200 volts, but a bias somewhat beyond plate-current cutoff, such as −250 volts, should be used. Therefore, since the allowable direct current through the bias resistor, r_1, of Fig. 33, is a maximum of 80 ma, a voltage drop of 250 volts requires a resistance of 3125 ohms for r_1. The power dissipated in r_1 is approximately 20 watts, and the power dissipated on the grid of the tube is comparatively small because of the low resistance of the grid when the grid current is large.

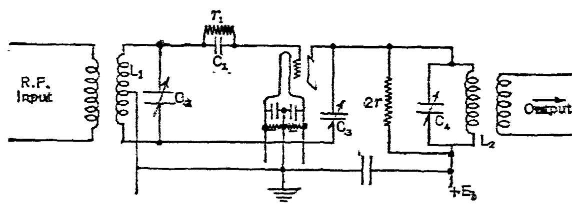


FIG. 33. Typical Circuit for Class C Amplifier

If it is assumed that one-third of the bias power is dissipated on the grid, the total power required of the r-f input circuit is approximately 30 watts.

This power must be supplied by the exciter amplifier, and the loss in the resistor r_1 cannot be supplied by means of a separate voltage supply in the hope that the exciter is required to supply the grid losses only. The grid-leak method of supplying bias as shown is a self-biasing arrangement and consumes no power from the plate supply. The one disadvantage of this method of supplying grid voltage is that if the excitation is lost the class C tubes have no bias, so that, in the case of the 204A, the resulting plate current would cause excessive plate dissipation. This difficulty may be removed by using a combination of a fixed bias (or biasing resistor in the cathode to negative plate supply lead) to limit the plate current if the excitation is lost and the additional bias supplied by a leak system as shown.

The input circuit to the tube is tuned and the inductance, L_1, is center-tapped in order that the amplifier may be neutralized by means of the condenser C_3. The plate circuit is tuned by means of C_4 and L_2, which in turn is coupled to an output circuit. The output circuit reflects, to the C_4L_2 circuit, an equivalent resistance of $2r$ across the tuned circuit. Therefore, r is the approximate resistance the tube works into during the time plate current flows.

POWER CALCULATIONS. The efficiency of the plate circuit of the class C amplifier can be estimated quite accurately if the peak a-c voltage across the tank circuit C_4L_2 is measured and used in eq. (17), as for the class B amplifier, provided the plate current is approximately one-half sine wave or that the fundamental component of the current is very large compared to the harmonic components. For this case it will be seen that the expression $I_{max}\,r$ is in reality the peak a-c voltage measured across the plate tank circuit in which I_{max} is the peak value of the fundamental component of plate current and r is the value of resistance the tube works into during the time plate current flows. The values of I_{max} and r are more or less fictitious, but their effective product is approximately the measured peak plate voltage.

Referring to the ratings on the 204A as given above, the permissible input plate power is 550 watts with only 167 watts plate dissipation, which requires a plate efficiency of approximately 70 per cent. Therefore, from eq. (17), it is found that the peak voltage across the tank circuit is 1778 volts. This peak voltage leaves only 222 volts on the plate to cause the peak plate current to flow. The power output for the above power input and plate loss is 383 watts, and if the peak plate voltage swing is 1778 volts, as calculated above, the approximate value of $2r$ is found to be 4150 ohms, which is the equivalent load resistance across the plate tank circuit, and the resistance the tube works into is r, or approximately 2075 ohms. The peak value of the fundamental component of plate current, as found from the expression for peak a-c plate voltage, is 860 ma. It is doubtful if this value of plate current flows at the instant the plate voltage reaches a minimum value of 222 volts. However, plate-circuit efficiencies of 70 per cent are easily attained, so that, even though the actual plate current may be considerably below 860 ma at the instant the plate voltage is 222 volts, the peak value of the fundamental component may be 860 ma without sufficient plate-current distortion to affect the plate loss seriously.

Experience has shown, however, in experimental class C amplifiers that approximately 70 to 75 per cent plate-circuit efficiency, for a circuit as shown in Fig. 33, is the limit of efficiency, because the actual plate current for higher load resistances deviates so far from a half sine wave of current at higher peak values of a-c plate voltage that the plate-current harmonic losses increase to keep the plate efficiency at a more or less constant value, in spite of the fact that the measured peak a-c plate voltage increases.

SPECIAL FILTER FOR HIGH EFFICIENCY. Equation (17) indicates that if an essentially half sine wave of plate current can be caused to flow to the plate the efficiency may be caused to increase as the peak tank voltage increases.

When the peak tank voltage approaches the plate voltage the efficiency by eq. (17) approaches 78.5 per cent. If the peak plate voltage could reach 1.27 times the d-c plate voltage, the efficiency would be 100 per cent, provided that the plate current is essentially a half sine wave and the plate resistance of the tube during the time current flows is zero. This 100 per cent efficiency condition is impossible, of course, but the limit of efficiency is seen to be 100 per cent, because the plate circuit is tuned and filters may be inserted in series with the plate to prevent appreciable flow of harmonic plate currents. Such a filter circuit is shown in Fig. 34. The inductance L_3 is comparatively large and offers a high impedance to the fundamental and harmonic plate currents. The primary function of L_3 is to provide a path for d-c plate current but to offer a high impedance to all a-c components of the plate current. The inductance L_4 and capacitance C_5 are tuned to series resonance for the fundamental frequency of the circuits. Therefore, the band-pass filter offers low impedance to the fundamental component of plate current but is designed to offer high impedance to harmonic components of the plate current.

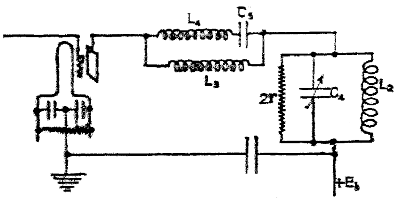


FIG. 34. Circuit for Increased Efficiency of Class C Amplifier

The value of the more or less fictitious plate resistance is a function of the tube and its excitation, and it may be calculated if all the constants of the tube and circuits are known. However, an apparent resistance of the tube may be easily calculated from the plate efficiency. This is done by assuming that the apparent load resistance is in series with an apparent plate resistance that will give the measured or calculated plate-circuit efficiency.

Such a formula is given below:

$$\frac{2r}{\text{Eff.}} - 2r = r_{ap} \text{ apparent plate resistance} \qquad (19)$$

in which $2r$ is the equivalent resistance in parallel with the tank circuit due to the load connected to the class C amplifier. Since the tube functions for approximately one-half the time, the actual effective resistance of the tube is approximately one-half of the apparent resistance r_{ap}, and the load the plate works into is one-half the equivalent load resistance or r. The load resistance r as calculated for the 204A at rated conditions as given above is 2075 ohms, and the calculated efficiency is 70 per cent. Substituting these values in formula (19), the effective resistance of the 204A is about 890 ohms.

If the harmonic components of the plate current are to be effectively reduced, the impedance of the circuit $L_3L_4C_5$, when tuned to the fundamental frequency, must be 4 or 5 times the approximate effective resistance of the tube at the second harmonic of the fundamental frequency. The impedance to higher harmonics, of course, will be still higher. Therefore, a rough approximation of what the impedance of the filter should be at the second harmonic is equal to the normal effective resistance across the output tank circuit at full power output of the tube.

Such impedance at the second harmonic frequency necessitates a comparatively large L to C ratio for the series tuned circuit L_4C_5. Since the a-c component of plate current flows through this series tuned circuit, the voltage across L_4 and C_5 reaches quite high values at the fundamental frequency, and the voltage across L_3 at all the harmonic frequencies is comparatively high. However, the system permits approximately only half sine waves of plate current at the fundamental frequency so that high efficiency can be achieved.

This type of filter has been successfully used to increase materially the efficiency of a class C amplifier in the laboratory, and was successfully applied to at least one broadcasting station, although the circuit has not been used extensively in transmitters to date. The effect of the filter on plate loss, power output, and efficiency of a laboratory test on a

204A is given in Fig. 35. It will be noted that for low load resistances, that is, resistances that limit peak a-c plate voltages to values appreciably below the d-c supply voltage, the filter adds little if any to the efficiency of the plate circuit. This is due to the fact that heavy loads cause essentially one-half sine waves of plate current, so that there are practically no harmonic components of plate current to reduce.

The results of the experimental data as plotted in Fig. 35 indicate that the efficiency of a class C amplifier without a filter is essentially constant over practically all the load range that might be used. The 204A tube, with which the data for the above curves were taken, was operated at approximately optimum grid excitation so that the values read from the curves represent approximately the best performance of the tube. Any loss in the filter for the curves in Fig. 35 is included as plate loss, so that the plate efficiency is actually greater than shown when the filter was used.

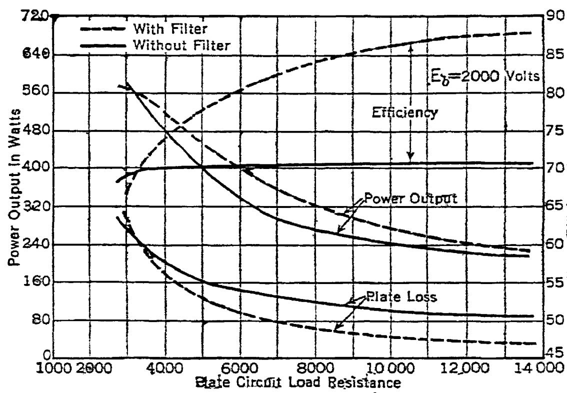


Fig. 35. Performance of RCA 204A as Class C Amplifier

Oscillograms of the actual plate current in a low-frequency class C amplifier were made on an experimental amplifier; they substantiate the above explanation of the behavior of the filter and indicate quite definitely that the plate current to the tube at most practical load resistances is a double peaked affair, whereas with the filter the plate current is essentially one-half of a sine wave. As a result of the peaked condition of the plate current without a filter, the peak emission required from the filament of the tube in a class C amplifier is 6 to 12 times the d-c value; when a filter is used the peak emission required is one-third to one-fifth of the peak without the filter. The reduced peak plate current requirement when the filter is used is a decided benefit.

Although the applications of the class C amplifier to radio devices are limited in number, it is the most efficient type of vacuum-tube amplifier known. When the class C amplifier is used as the output system for a radiophone transmitter, in which a class B audio amplifier is used to modulate the plate supply to the class C amplifier, a radiophone transmitting station is obtained that has a greater overall efficiency than any other type of transmitter now in general use.

SUMMARY FOR CLASS C AMPLIFIERS. Some of the more important circuit requirements of the class C amplifier as well as some of the more important operating points are summarized below:

1. If full output is to be obtained from the class C amplifier, the grid excitation should be sufficient to obtain essentially full permissible d-c grid current at the recommended bias.

2. The grid excitation for a plate-modulated class C amplifier must be sufficient to cause a linear relation between the plate voltage and output voltage.

3. In a plate-modulated class C amplifier, it must be remembered that the peak plate voltage due to modulation peaks is double the plate voltage for normal carrier, which in turn doubles the average plate current and also the peak current corresponding to a power four times the carrier power. Therefore, it is necessary to determine whether the tubes used in such an amplifier are capable of withstanding the peak d-c plate voltage over an audio cycle and whether the emission of the tube is ample to meet the peak-plate-current conditions.

4. Failure to meet the peak-plate-current condition in item 3, because of emission limitation or insufficient excitation, results in a second harmonic in the detected carrier output. Failure to meet the requirements in item 3 only reduces the power output and efficiency in the case of class C amplifiers for telegraphy.

5. The bias for a class C amplifier must be beyond plate-current cutoff at the highest instantaneous d-c plate voltage. For 100 per cent modulation the bias should be 2.5 to 3 times the bias for plate current cutoff at normal plate voltage. Excessive bias usually results in higher excitation power with little gain in efficiency.

6. The plate dissipation of a tube in a class C amplifier is determined entirely by the plate supply voltage and the load resistance for normal class C operation of the tube, and the value of load resistance has essentially no relation to the plate resistance of the tube.

7. The maximum plate efficiency of a class C amplifier is approximately 70 to 75 per cent for all practical load resistances. This efficiency may be raised to 80 or 90 per cent by using somewhat higher resistance plate loads if a band-pass filter is used in series with the plate.

8. The band-pass filter also reduces the peak-plate-current requirements, which tends to reduce distortion at high percentage modulation.

9. The audio power required to modulate the plate circuit is 50 per cent of the amplifier input plate power; therefore, any increase in plate efficiency reduces the modulator power required.

4. REGENERATION AND ITS PREVENTION

CONDITIONS FOR REGENERATION. In each of the theoretical amplification equations (eqs. [4], [6], etc.) the internal input impedance z_{i2} appears. When r_{i2} is a positive quantity the greatest amplification will be obtained by making the internal input impedance as great as possible. However, it is possible for r_{i2} to have a negative value (p. 5-49) when the external plate impedance of the succeeding tube is inductive. In this case the conditions under which maximum amplification will be obtained are quite different.

If r_i is negative and less than twice the magnitude of the resistance presented to the grid filament terminals by the rest of the input circuit, it is possible for the amplification to be greater than the value obtained when the internal input impedance is infinite, for the plate current of a single tube in terms of the impressed voltage is (see eq. [1], p. 7-03)

$$i_{pa} = \frac{\mu e'_g(r_i + jx_i)}{r_i + r_e + j(x_i + x_e)}\left[\times \frac{z}{r_p + z}\right] \tag{20}$$

and the absolute value is

$$i_{pa} = \frac{\mu e'_g\sqrt{r_i^2 + x_i^2}}{\sqrt{(r_i + r_e)^2 + (x_i + x_e)^2}}\,\frac{\sqrt{r^2 + x^2}}{\sqrt{(r + r_p)^2 + x^2}} \tag{20a}$$

If either r_i or x_i is infinite, the first factor reduces to $\mu e'_g$, but if r_i is negative and less than twice the magnitude of r_e (and x_i and x_e are of opposite sign and their sum less than x_i) the first factor will be greater than $\mu e'_g$. The tube is then said to be *regenerating*. In particular when the denominator is zero the amplification will be infinite and the tube will *oscillate;* that is, plate current of a particular frequency will flow even when there is no externally impressed alternating voltage on the grid.

In an impedance-capacitance coupled amplifier (eq. [6], p. 7-04) regeneration occurs when

$$0 > r_{i2}/z_{i2}^2 > 2(r/z^2 + 1/r_p)$$

When transformer-coupled amplification is used and the circuit is properly tuned (eq. [7b]) regeneration occurs when $0 > r_{i2} > 2r_4$, and oscillation when $r_{i2} = -r_4$.

The *equivalent negative resistance* of the input circuit is caused by the plate voltage electrostatically inducing a voltage between the grid and filament which has a component in phase with the impressed grid voltage. This will occur only when the plate load is inductive. (See p. 5-49.) A similar voltage impressed on the grid by any other means would, of course, have the same effect. (See Regeneration, Theory and Experiment, by Peterson, Kreer, and Ware, *Proc. I.R.E.*, October, 1935.) Circuits are frequently designed so that part of the energy in the plate circuit is "fed back" to the grid circuit to increase the amplification. When the feedback is inductive through a movable coil, this coil is called the *tickler coil*.

EFFECTS OF REGENERATION. The increased amplification obtained as a result of regeneration is sometimes used to increase gain, particularly where a single frequency is being amplified. Since a regenerative circuit usually behaves like a very sharply tuned circuit it is not useful when a broad band of frequencies must be amplified.

Regeneration results in a decrease in the effective input resistance of the amplifier and so decreases the non-regenerative amplification obtainable from the tube.

As a rule most difficulty is experienced at radio frequencies. Even if regeneration is prevented at the working frequency of the amplifier, there may be considerable feedback at higher frequencies: if a resonant path exists at one of these high frequencies a *parasitic self-oscillation* may begin. This may cause the amplifier to cease functioning as an amplifier, and unless some protective device limits the plate current the tubes may be overheated very quickly.

Another type of parasitic oscillation may exist in the pushpull circuit shown in Fig. 31; in which the input circuit has a certain impedance from ground to grids, because of the leakage in the center-tapped grid and plate inductances. This condition effectively places the pushpull tubes in parallel with an input and output circuit with the neutralizing condensers ineffective. If the input and output circuits have sufficiently high impedance (which may be comparatively low for high-power tubes) parasitic oscillations will result. However, it should be noted that r_1 and r_2 in Fig. 31 are connected from their respective grids to ground so that the resistors are still effective in suppressing parallel parasitic oscillation of the tubes. If the corresponding value of resistance in r_1 and r_2 is placed across the grids without the r-f ground at the center point, parallel parasitic oscillation of the tubes may exist because the grids would be at essentially the same potential so that the resistances would be ineffective.

This type of parasitic oscillation is apparently responsible for the self-oscillation of most pushpull audio amplifiers.

PREVENTION OF OSCILLATION. Regeneration to the point of oscillation may be prevented in several ways. It is often accomplished by the introduction of power-absorbing resistors into various parts of the amplifier circuit, but these invariably increase the losses in the circuit and so are objectionable. Another possible method is to decrease the impedance of the input circuit, but this is frequently impracticable without too great decrease in amplification.

Parasitic oscillations may sometimes be prevented by so changing either the plate or grid circuit that the resonant frequencies of the one have no counterpart in the other. There is then no complete resonant path at any frequency. In general, if the leads to the grid are short, and a small r-f choke is inserted in series with the plate to increase the effective length of the plate leads, there will be no common resonant frequency. If the circuits cannot be so unbalanced resistances must be used, such as are shown in Fig. 31.

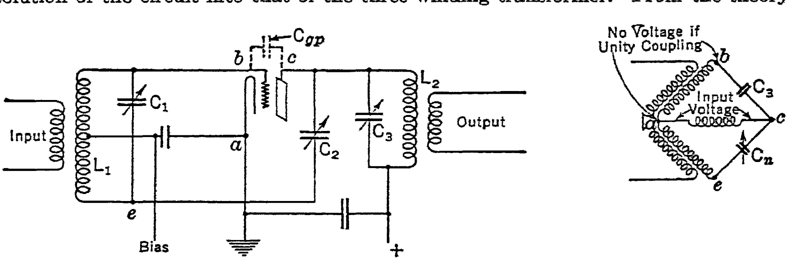
The most effective cure for parasitic oscillation of pushpull audio amplifiers is to unbalance the input and output circuits by placing a resistor or small condenser across one side of the input or output transformers. In general, the value of impedance required to suppress the undesired oscillation will not noticeably affect the frequency characteristic of the amplifier.

Improvement in shielding between stages will sometimes eliminate enough of the feedback to make the circuit stable, or the plate and grid circuits may be moved apart if practicable. Shielding for pentode or tetrode r-f amplifiers is always necessary. Shielding an audio amplifier is usually not necessary with proper placement of the various tubes and transformers for minimum feedback. However, it may be necessary to use elaborate means to prevent stray pick-up to the input transformer. In general, a magnetic shield for audio frequencies is better than a copper shield, but the copper is better at radio frequencies.

Where the feedback is due to the grid-plate capacitance, this capacitance may be eliminated by electrostatically shielding the plate from the grid. This has been done in the tetrode and pentode. It makes possible the extremely high amplification factors obtainable with these tubes.

NEUTRALIZATION. When it is impossible to eliminate the fed-back voltage it is still possible to cancel its effect by introducing an equal and opposite voltage. Circuits to accomplish this are based on the impedance bridge principle or on the three-winding transformer. (See p. 6-12.)

The first method for *neutralizing* the grid-plate capacitance was that of Rice shown in Fig. 36. Diagram A of this figure is a schematic sketch of connections; diagram B is the resolution of the circuit into that of the three-winding transformer. From the theory of

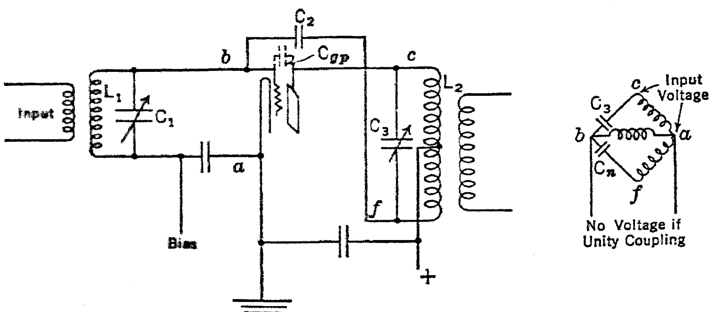


A. Schematic Diagram *B*. Equivalent Circuit

FIG. 36. Rice System for Neutralization. The tuning and bypass capacitors shown in the circuit diagram on the left are omitted from the equivalent three-winding transformer diagram on the right.

the three-winding transformer any voltage induced in the coil *a-c* will cause no voltage drop across *b-a-e* (except that due to incidental resistance of the coils) if the value of C_n is properly chosen and the coupling between *b-a* and *a-e* is unity. In practice the condenser (C_n) is made variable to permit its proper adjustment and is called the neutralizing condenser. The disadvantages of the Rice method lie in the fact that part of the input voltage across *a-e* is lost since it is not impressed on the grid, and that neither side of the neutralizing condenser can be grounded.

A modification of the Rice method is shown in Fig. 37, *A* being a schematic diagram of connections. Diagram *B* shows that this circuit can also be resolved into a three-winding transformer circuit. From the theory there will be a value of C_n for which the voltage across *b-a* will be very low, approaching zero as the resistance of *c-a-f* decreases. This system is not so useful for power amplifiers as the Rice system; the plate circuit of the latter can be more favorably loaded because the entire plate coil is in the plate circuit.



A. Schematic Diagram *B*. Equivalent Circuit

FIG. 37. Modified Rice System for Neutralization. The tuning and bypass capacitors shown in the circuit diagram on the left are omitted from the equivalent three-winding diagram on the right.

Both these methods require that for perfect neutralization the coefficient of coupling be unity and that the incidental resistances be zero; hence perfect neutralization can be only approached.

A pushpull r-f amplifier can be neutralized, as shown in Fig. 31, by means of the condensers C_6 and C_7, the value of which is equal to the grid-plate capacitance of the tubes for symmetrical input and output circuits.

In general, shielding of the input and output circuits of a capacitance-neutralized amplifier need not be so complete or effective to prevent feedback as for the screen-grid tube amplifier, because some stray magnetic or capacitance feedback can be effectively compensated for by the proper adjustment of the neutralizing condensers for minimum feedback. A common and convenient method to adjust the neutralizing capacitance properly is to apply a signal to the grid of the amplifier to be neutralized with the tube in place and the filament lighted but without plate voltage. An r-f galvanometer is placed in the plate tank circuit, and the neutralizing capacitances are adjusted for zero or minimum galvanometer deflection. Precaution against excessive currents through the galvanometer must be taken for the initial adjustment of the neutralizing condensers, or other methods of indicating a tank voltage may be used.

If the amplifier to be neutralized is a plate-modulated class C amplifier the simplest and quickest method for neutralizing is to overmodulate the amplifier and adjust the neutralizing condenser until the instantaneous value of carrier will be zero at the peaks of downward modulation. A cathode-ray oscillograph may be used as an indicator for the adjustment.

BIBLIOGRAPHY

Barton, Loy E., High Audio Power from Relatively Small Tubes, *Proc. I.R.E.*, July, 1931.

Fay, C. E., The Operation of Vacuum Tubes as Class B and Class C Amplifiers, *Proc. I.R.E.*, March, 1932.

Thompson, B. J., Graphical Determination of Performance of Push-pull Audio Amplifiers, *Proc. I.R.E.*, April, 1933.

Warner, J. C., and A. V. Loughren, The Output Characteristics of Amplifier Tubes, *Proc. I.R.E.*, December, 1926.

Other articles relating to vacuum-tube amplifiers may be found by referring to the *Proc. I.R.E.* General Index and to the index of *J. A.I.E.E.*, *Bell System Tech. J.*, Bureau of Standards Circulars, *Gen. Elec. Rev.*, *Phys. Rev.*, *QST*, and other periodicals on radio and electrical subjects.

SPECIAL-PURPOSE AMPLIFIERS

By E. L. Clark

A special-purpose amplifier, as the name implies, is a particular type of vacuum-tube amplifier designed for a definite purpose. Such amplifiers are usually of the capacitance, impedance-coupled circuit type, as differentiated from tuned and mutually coupled amplifiers.

A wide-band amplifier is an amplifier that covers a wide frequency spectrum, such as is used for a television video amplifier. In most cases, the wide-band amplifier operates as a class A amplifier with plate current flowing for 360 electrical degrees. A wide-band amplifier is characterized by low gain; for a given tube the gain multiplied by the band width is a constant.

A cathode follower is an amplifier which has its output load in the cathode circuit of a vacuum tube. The characteristics of a cathode follower are: reduced input capacitance, increased input resistance, reduced output impedance, a gain of less than 1, and no voltage inversion. The cathode follower may be used as an impedance matching device, without the use of a transformer.

A grounded-grid amplifier is an amplifier that has its grid grounded, and the input signal applied to the cathode. The output circuit is in the plate, in the usual manner. The characteristics of a grounded-grid amplifier are: low input impedance, low input-to-output capacitance due to the shielding action of the grounded grid, and no voltage inversion.

An in-phase amplifier is one in which the polarity of the output signal is the same as that of the input signal. This is of no importance in sine-wave work. However, for pulse work and for television video amplifiers, the polarity of the amplified signal is of great importance.

A negative feedback amplifier is an amplifier in which some of the output signal is fed back to the input to modify the output. There are two general types of feedback amplifiers, the voltage-feedback and the current-feedback types. Both are characterized by a reduction in gain. However, the voltage-feedback type gives an apparent decrease in plate resistance. Both types of feedback give a reduction in distortion.

A one-shot amplifier is an amplifier which, after responding to an input pulse, will not respond to a second pulse until a given time has elapsed. This amplifier is similar to a multivibrator, and it requires two tubes. The characteristic of a one-shot amplifier is that, after a given input level has been reached, the output is a sudden sharp pulse which cannot be immediately repeated, a definite time interval being required before a second pulse can be obtained.

A pulse amplifier is an amplifier that is designed to handle a pulse type of input signal. Pulse signals are of two general types: positive pulses and negative pulses. The pulse amplifier must be designed for the polarity of pulse to obtain the optimum performance. The band width of the pulse amplifier must also be adjusted to the type of pulse being handled if the pulse shape is not to be degraded and if maximum gain is to be obtained.

The characteristic of a pulse amplifier intended to handle positive pulses requires that it be biased nearly to cutoff and that it shall approach class B operation and efficiency. However, to handle negative pulses the vacuum-tube bias must be near zero, and a heavy plate current is drawn except when the pulse signal is applied. This results in low efficiency for a negative pulse amplifier.

5. WIDE-BAND AMPLIFIERS

The wide-band amplifier is used when the frequency response must be extended beyond about 15 kc. It finds its chief application as a video amplifier in television equipment. However, it is finding other uses such as in radar and pulse work. In order better to understand the frequency response limitations of an r-c coupled amplifier, see Fig. 1. This is a curve of output voltage vs. frequency, of a typical amplifier stage, as shown in Fig. 2. The frequencies $5f_1$ and f_2 are considered the useful limits of the wide-band amplifier. This is somewhat empirical, but practice has shown it to be desirable in order to reduce the low-frequency phase shift (see p. 7-46). If many cascaded stages are used it may be desirable to take $10f_1$ instead of $5f_1$ as the low-frequency limit of the amplifier.

HIGH-FREQUENCY RESPONSE. The high-frequency range will be defined as that region lying above about 1 kc. The simplest form of wide-band amplifier is that shown in Fig. 2 with a low value of plate resistor (r_L). In general, pentodes are used for wide-band amplifiers, for two reasons. First, the Miller capacitance effect is negligible; second, they are made with a higher mutual conductance (g_m) than triodes. The general formulas for a class A amplifier, eqs. (1), (1a), and (2) (p. 7-03), also cover wide-band amplifiers. However, the value of Z is of interest and governs the gain and band width of the amplifier. In Fig. 2, C_T is the total shunting capacitance, composed of the stray circuit capacitance, the output capacitance of V_1, and the input capacitance of V_2 (which may be partly composed of Miller capacitance in triodes). Let the resistance r_0 represent the resulting

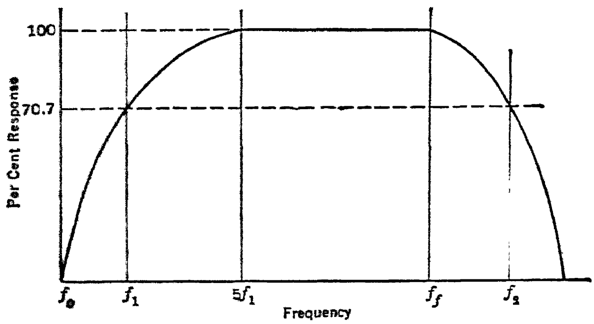


FIG. 1. Amplifier Response Curve

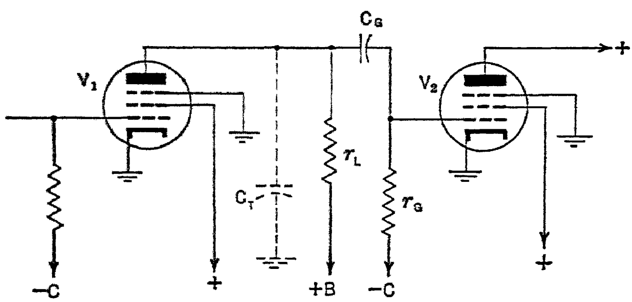


FIG. 2. Uncompensated Amplifier Stage

parallel resistance of the plate resistor r_L and the grid resistor r_g. The absolute value of Z is given by

$$Z = \frac{r_0 x_{C_T}}{\sqrt{r_0^2 + x_{C_T}^2}} \tag{1}$$

At the frequency f_2 which makes the capacitance reactance x_{ct} equal to r_0, the response will be down to 70.7 per cent of the maximum response. This gives a means of determining the value of C_T.

$$C_T = \frac{1}{2\pi f_2 r_0} \tag{2}$$

As the plate resistor r_L is usually much smaller than r_g, the effect of r_g is usually negligible.

Using g_m, the mutual conductance of the vacuum tube, the gain of the stage is

$$\text{Gain } A = \frac{g_m r_p Z}{r_p + Z} \tag{3}$$

However, for wide-band amplifiers using pentodes, r_p is much greater than Z, therefore r_p can usually be disregarded, and the gain then becomes

$$\text{Gain } A = g_m Z \tag{4}$$

The maximum gain of the flat portion of the curve of Fig. 1 is simply $g_m r_L$.

The phase angle (ϕ) is

$$\phi = \tan^{-1} \frac{r_L}{x_{C_T}} \tag{5}$$

for parallel resistance and capacitance as shown in Fig. 2 which represents a time delay of

$$\frac{0.035}{f_2} \text{ seconds} \tag{6}$$

For an uncompensated amplifier: Fig. 3 gives the value of r_L for various values of shunt capacitance C_T, at the frequency f_2. The response will be down to 70.7 per cent of the flat portion of the curve at frequency f_2 when using a plate load r_L as determined from Fig. 3.

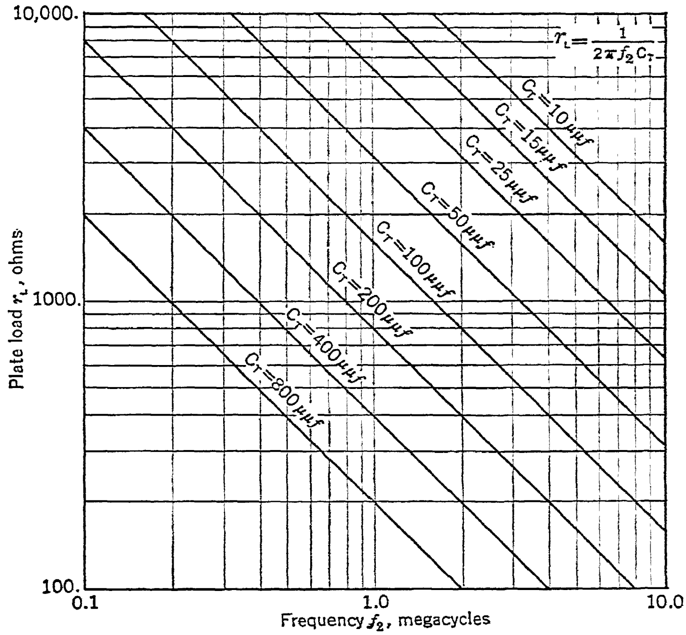


FIG. 3. Plate Load (r_L) in Terms of Frequency (f_2) and Total Capacitance (C_T) for an Uncompensated or Shunt Peaked Amplifier

SHUNT PEAKING. In Fig. 4 is shown the schematic of an amplifier stage using shunt peaking. As previously described, C_t is the total shunting capacitance. The gain for this amplifier is given by eq. (4). The value of Z is

$$Z = X_{C_T}\sqrt{\frac{r_L^2 + x_L^2}{r_L^2 + (x_L - x_{C_T})^2}} \tag{7}$$

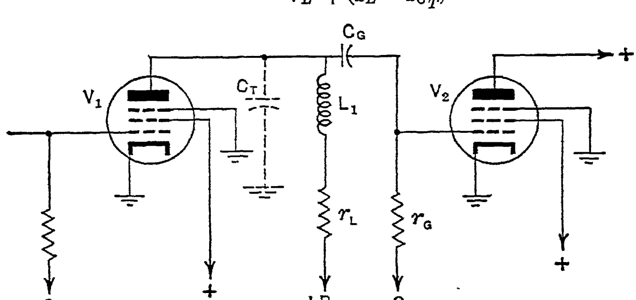


FIG. 4. Circuit for a Shunt Peaked Amplifier Stage

If an inductance L_1 is chosen so that $x_{L1} = \frac{1}{2} x_{C_T}$ at f_2, the resulting impedance Z will be equal to r_L. This means that the response will be flat to frequency f_2 instead of being

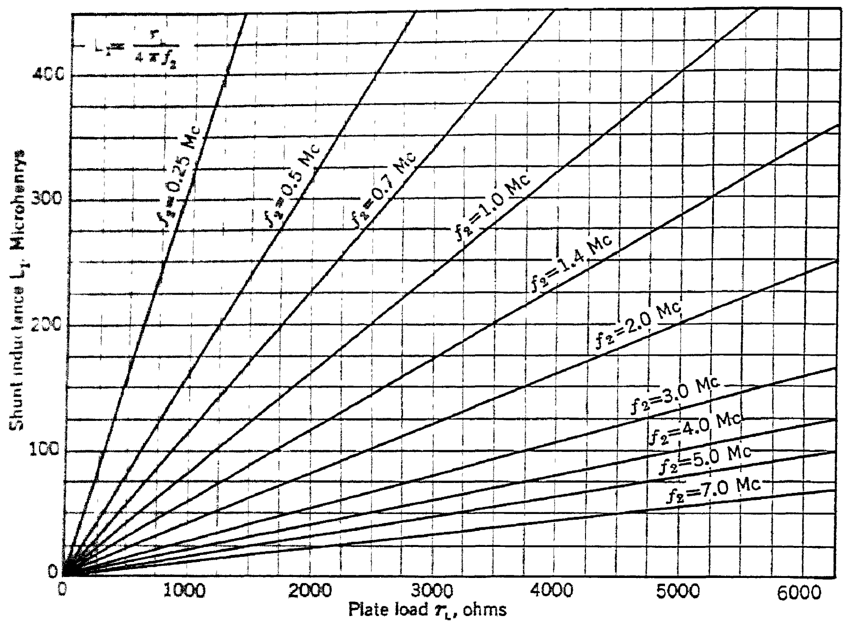


FIG. 5. Shunt Inductance (L_1) in Terms of Frequency (f_2) and Plate Load (r_L) for a Shunt Peaked Amplifier

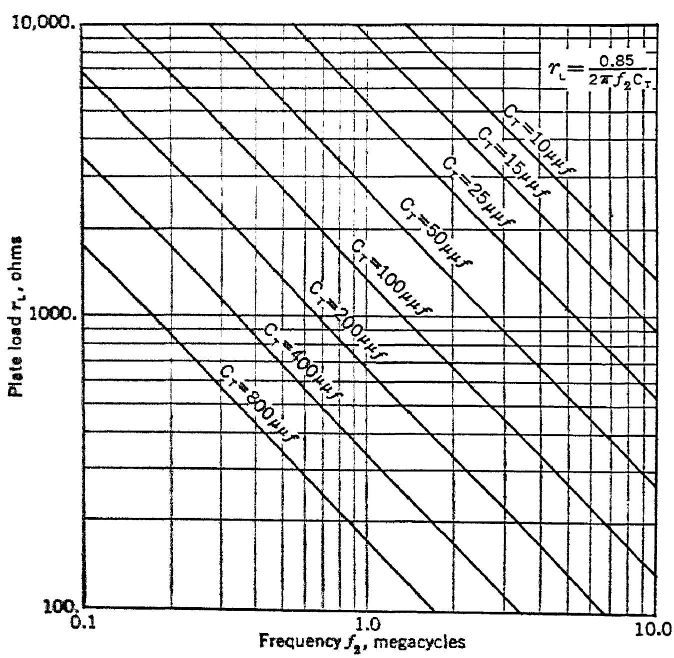


FIG. 6. Plate Load (r_L) in Terms of Frequency (f_2) and Total Capacitance (C_T) for a Shunt Peaked Amplifier, with Corrected Phase and Amplitude Characteristics

down to 70.7 per cent as in the uncompensated amplifier. The values are

$$r_L = \frac{1}{2\pi f_2 C_t} \tag{8}$$

$$L_1 = \frac{r_L}{4\pi f_2} \tag{9}$$

$$\text{Gain } A = g_m r_L \tag{10}$$

The time delay is no greater than $0.023/f_2$ seconds.

Figure 5 gives the value of inductance L_1, for various frequencies f_2, and load resistance r_L for use in a shunt peaked amplifier.

A shunt peaked amplifier can be designed quickly by the use of Fig. 3 and Fig. 5. Knowing what frequency f_2 is needed (say 3.0 megacycles), determine the value of C_t; 40 $\mu\mu f$ is

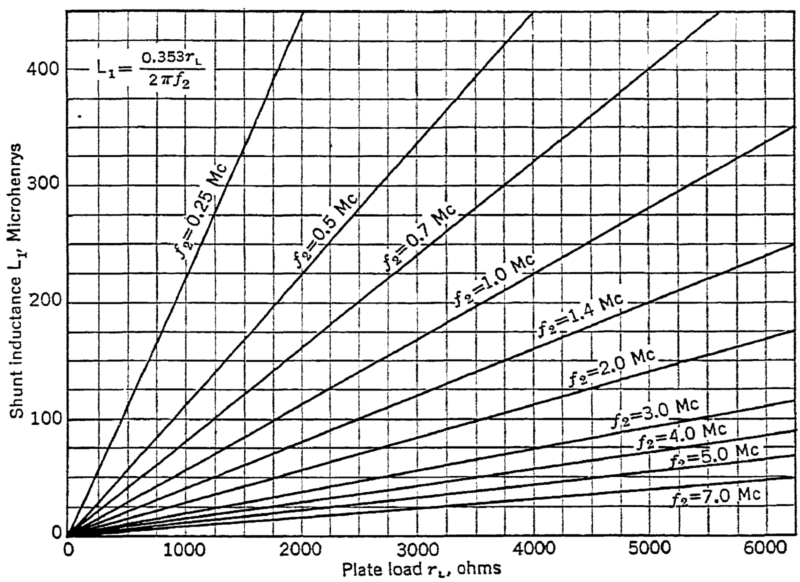


FIG. 7. Shunt Inductance (L_1) in Terms of Frequency (f_2) and Plate Load Resistance (r_L) for a Shunt Peaked Amplifier with Corrected Phase and Amplitude Characteristics

a good estimate. Then from Fig. 3 find r_L to be 1330 ohms. Now from Fig. 5 at 3 megacycles and 1330 ohms, L_1 is found to be 35.2 microhenrys. The stage gain will be 1330 times the mutual conductance g_m of the tube in mhos.

For multistage wide-band amplifiers where the best phase and amplitude characteristics are needed, slight revisions in eqs. (8) and (9), as shown by Freeman and Schantz, will give almost perfect results up to frequency f_2

$$r_L = \frac{0.85}{2\pi f_2 C_t} \tag{11}$$

$$L_1 = \frac{0.353 r_L}{2\pi f_2} \tag{12}$$

Equations (11) and (12) are plotted in Fig. 6 and Fig. 7 to facilitate the design of such an amplifier.

SERIES PEAKING. Figure 8 is the schematic diagram of a series peaked amplifier. It can be seen from Fig. 8 that the capacitance is split by the inductance L_2. This results in the vacuum tube V_1 working into a smaller capacitance than in the previous case of shunt peaking, with the results that more gain is obtained. For proper operation the ratio of $C_1/C_2 = 1/2$. If $C_1/C_2 = 2$ the plate load resistor r_L must be put on the other side of the series inductance L_2. The rule is to keep the plate load resistance r_L on the low-

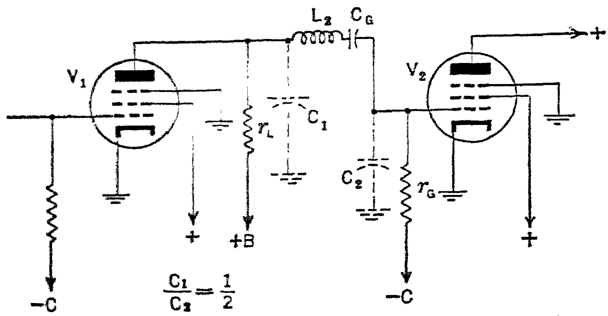


FIG. 8. Circuit for a Series Peaked Amplifier when $C_1/C_2 = 1/2$

capacitance side of the series inductance L_2. The value of r_L for series peaking is given by

$$r_L = \frac{1}{4\pi C_1 f_2} \tag{13}$$

where $C_1/C_2 = 1/2$. In most cases, however, it is more convenient to use C_T, the sum of C_1, C_2, and strays, since C_T can be determined more accurately than C_1. In terms of C_T, r_L is given by

$$r_L = \frac{1.5}{2\pi f_2 C_T} \tag{14}$$

If the series inductance L_2 is chosen to resonate with C_1 at a frequency of $f_2\sqrt{2}$, the response will be flat to frequency f_2. The value of the series inductance L_2 is given by

$$L_2 = \frac{1}{8\pi^2 f_2^2 C_1} \tag{15}$$

or, in terms of C_T, $L_2 = 667 C_T r_L^2$.

Substituting the value of C_T from eq. (14)

$$L_2 = \frac{r_L}{2\pi f_2} \tag{16}$$

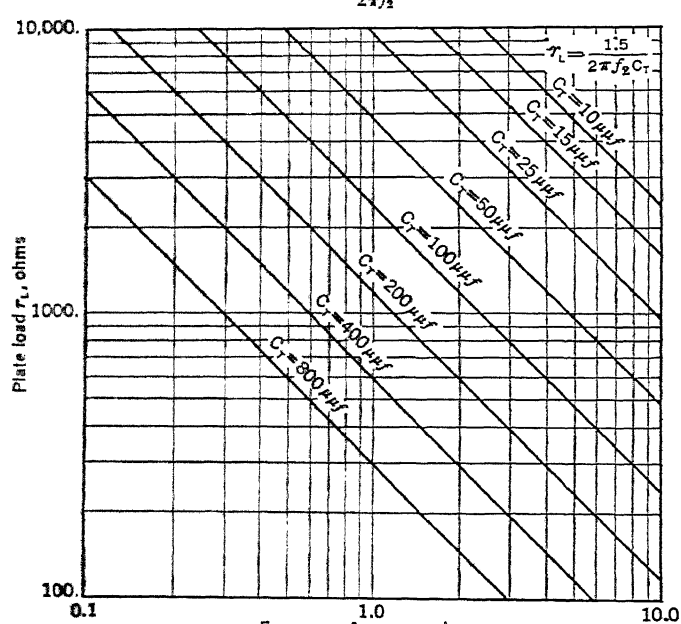


FIG. 9. Plate Load (r_L) in Terms of Frequency (f_2) and Total Capacitance (C_T) for a Series Peaked Amplifier Stage

The phase delay is a rather complicated function in the series peaking circuit, but in general, as Seeley and Kimball have shown, the time delay up to the frequency f_2 is constant within a variation of $0.0113/f_2$ seconds. This is roughly one-half the variation in

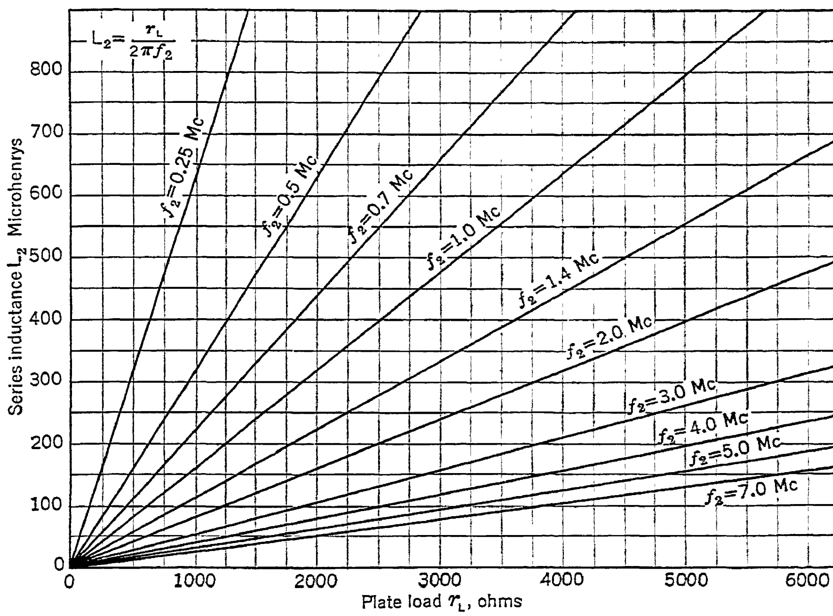


FIG. 10. Series Inductance (L_2) in Terms of Frequency (f_2) and Plate Load (r_L) for a Series Peaked Amplifier Stage

phase delay experienced with shunt peaking. The gain of a series peaked amplifier is given by eq. (10), using the value of r_L as obtained from eq. (14).

To facilitate the design of a series peaked amplifier eq. (14) is plotted in Fig. 9 and eq. (16) is plotted in Fig. 10. The curves of Fig. 9 and 10 are to be used as described in *shunt peaking*.

COMBINATION OF SHUNT AND SERIES PEAKING. As might be expected, the advantages of shunt and series peaking can be combined to increase the gain further.

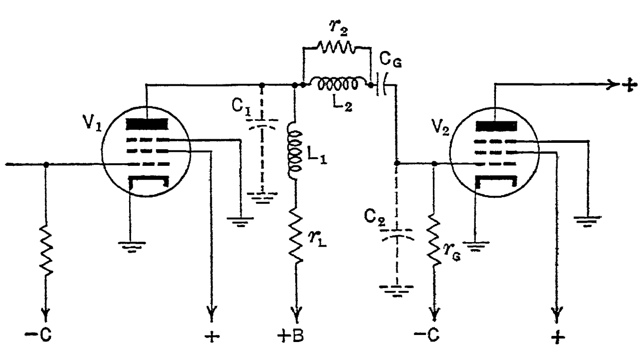


FIG. 11. Circuit for a Combined Shunt and Series Peaked Amplifier Stage

The series inductance L_2 separates the output capacitance C_1 and the input capacitance C_2, while the shunt inductance L_1 compensates for the output capacitance C_1. The circuit is shown in Fig. 11.

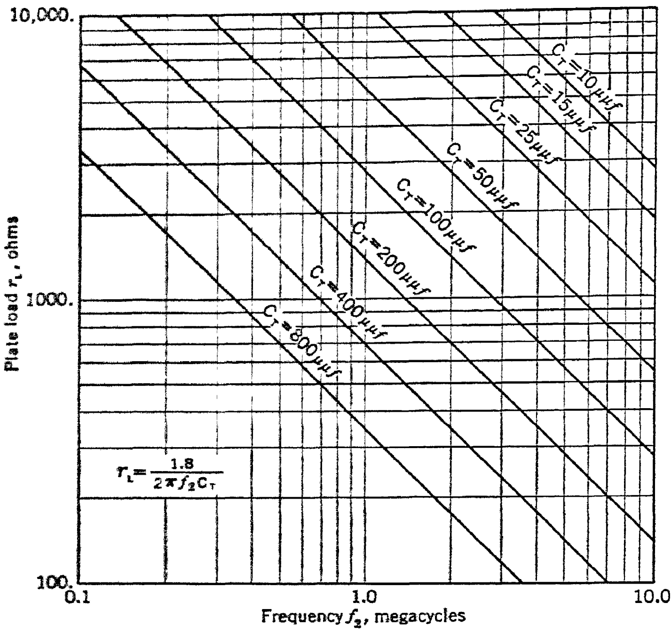


FIG. 12. **Plate Load** (r_L) in Terms of Frequency (f_2) and Total Capacitanec (C_T) for a Combined Shunt-series Peaked Amplifier Stage

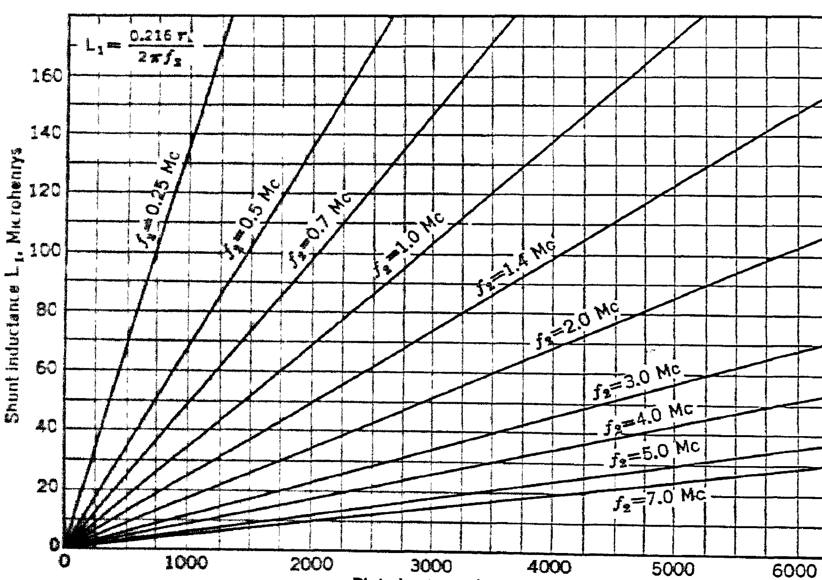


FIG. 13. **Shunt Inductance** (L_1) in Terms of Frequency (f_2) and Plate Load Resistance (r_L) for a Combined Shunt-series Peaked Amplifier Stage

Again, as in series peaking, the ratio of $C_1/C_2 = {}^1/_2$ for proper operation.

$$r_L = \frac{1.8}{2\pi f_2 C_T} \tag{17}$$

$$L_1 = 0.12 C_T r_L^2 \tag{18}$$

$$L_2 = 0.52 C_T r_L^2 \tag{19}$$

If the value of C_T from eq. (17) is substituted in eq. (18) and (19), they become

$$L_1 = \frac{0.216 r_L}{2\pi f_2} \tag{20}$$

$$L_2 = \frac{0.936 r_L}{2\pi f_2} \tag{21}$$

The gain of a combined peaked amplifier is given by eq. (10) using the value of r_L as obtained from eq. (17).

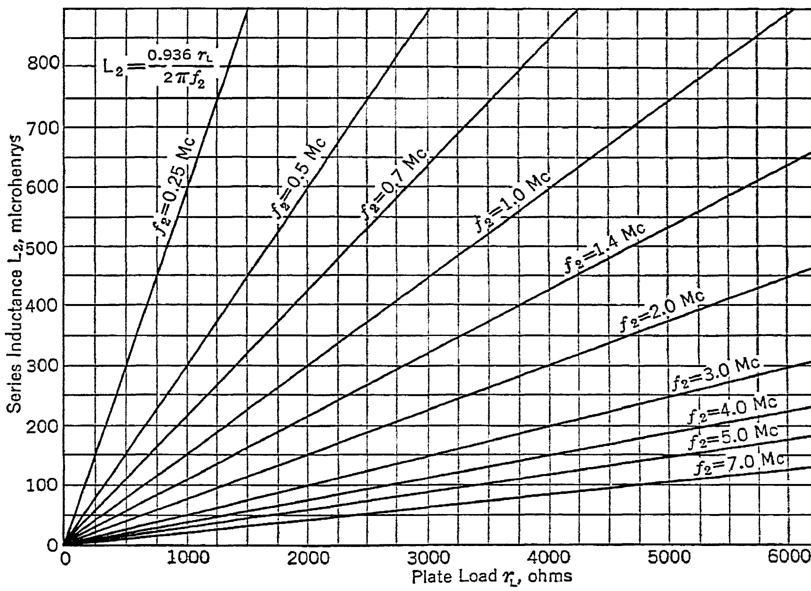


FIG. 14. Series Inductance (L_2) in Terms of Frequency (f_2) and Plate Load Resistance (r_L) for a Combined Shunt-series Peaked Amplifier Stage

The phase-delay expression becomes more complicated in this type of peaking, but it does not exceed $0.015/f_2$ seconds, up to f_2.

If the Q of L_2 is too high, a high-frequency peak will be experienced just before f_2 is reached. The resistance r_2 shunting L_2 is to lower the Q of the inductance L_2 and prevent the formation of a peak in the response curve. The value of resistance r_2 may vary from 5 to 10 times the load resistance r_L.

Equations (17), (20), and (21) are plotted in curves, Figs. 12, 13, and 14, to expedite the design of a combined shunt-series peaked amplifier.

CONSTANT-*K*-TYPE FILTER COUPLING NETWORK. The schematic diagram, Fig. 15, shows a wide-band amplifier of the constant-K, low-pass filter type. This circuit appears at first glance like the combined shunt-series peaking network; however, its constants are based on standard constant-K low-pass filter equations, as follows: $L_2 = \dfrac{r}{\pi f_2}$ and $C_2 = \dfrac{1}{\pi f_2 r}$. In forms to apply to the circuit of Fig. 15 these become:

$$r_L = \frac{1}{\pi f_2 C_2} \tag{22}$$

$$L_2 = r_L^2 C_2 = \frac{r_L}{\pi f_2} \tag{23}$$

$$L_1 = \frac{L_2}{2} \tag{24}$$

and r_2 = 5 to 10 times r_L. The stage gain is still given by eq. (10) using the value of r_L from eq. (22). This indicates higher gain than any of the other high peaking systems. This is true; however, one factor has been ignored in all the foregoing systems, namely, distributed capacitance. Only the distributed capacitance C_d across the series coil as

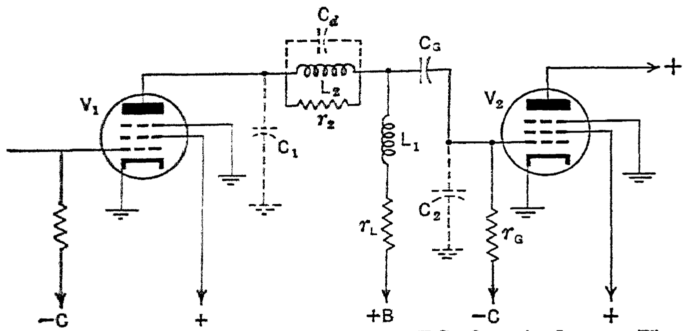


FIG. 15. Circuit for a Wide-band Amplifier with a Constant-K Configuration Low-pass Filter-coupling Network

shown in Fig. 15 has much effect. This capacitance, however, changes the seeming constant-K low-pass filter coupling network into an M-derived filter section. In most wide-band amplifiers the frequencies encountered are high enough so that this effect cannot be ignored. In changing from a constant-K to an M-derived low-pass filter, there are two ways to keep a given pass band: to reduce the shunting capacitance, or to reduce the characteristic impedance. As the capacitance cannot be reduced, the only course is to reduce the load resistance r_L, thus lowering the gain. The equations, taking this distributed capacitance C_d into consideration, are as follows:

$$M = \sqrt{1 + \left(\frac{C_d}{C_1}\right)^2} - \frac{C_d}{C_1} \tag{25}$$

$$r_L = \frac{M}{\pi f_2 C_2} \tag{26}$$

$$L_2 = r_L{}^2 C_2 = \frac{M r_L}{\pi f_2} \tag{27}$$

Also $L_1 = 0.8L_2$ approximately, and r_2 = 5 to 10 times r_L.

Equation (28) gives the frequency of infinite attenuation which should be kept well outside the pass band to prevent excessive phase shift.

$$f\infty = \frac{1}{\pi\sqrt{2L_2C_1(1 - M^2)}} \tag{28}$$

Under normal conditions the phase shift is about the same as for the combined shunt and series peaking system.

Equation (25) is plotted in Fig. 16, which gives the values of M in terms of the output

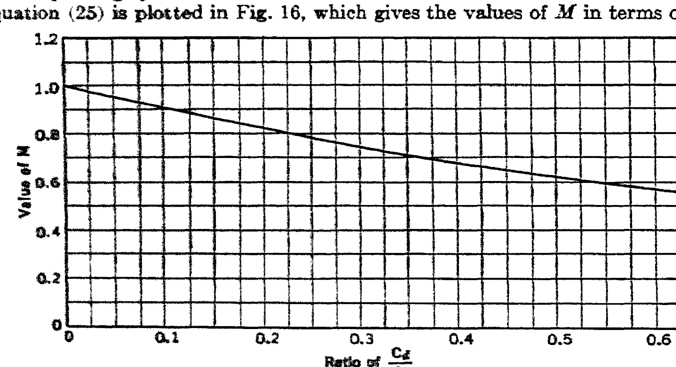


FIG. 16. Value of M as Produced by the Distributed Capacitance (C_d) of the Series Peaking Coil (L_2) and the Output Capacitance (C_1) of V_1 (Fig. 15)

capacitance C_1 and the distributed capacitance C_d of the series inductance L_2. The curve in Fig. 17 shows the effect of M on the load resistance r_L and the series inductance L_2. As the stage gain is directly proportional to the load resistance r_L, the importance of keeping the distributed capacitance C_d as low as possible is apparent.

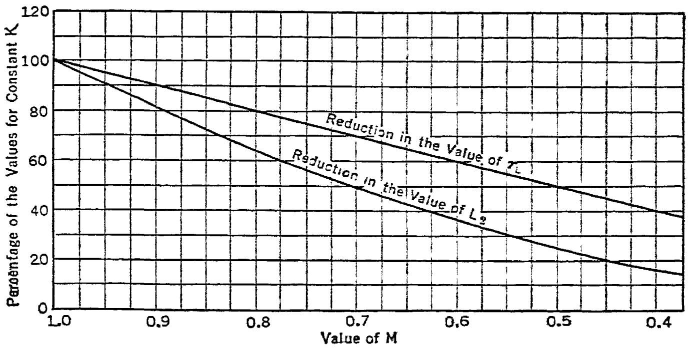


FIG. 17. The Effect of M on the Plate Load (r_L) and the Series Peaking Inductance (L_2)

Equations (22), (23), and (24) are plotted in the form of curves in Figs. 18 and 19 to be used for designing wide-band amplifiers of the constant-K configuration. Knowing the top frequency f_2 (say 3 megacycles), determine the value of C_2; 25 $\mu\mu$f is a good estimate. Then from Fig. 18 find r_L to be 4240 ohms. Now from Fig. 19 at 3 megacycles and 4240 ohms find L_2 = 448 microhenrys, and L_1 = 224. However, the distributed capacitance of L_2 must be considered. It may be about 4 $\mu\mu$f while C_1 may be approximately 16 $\mu\mu$f.

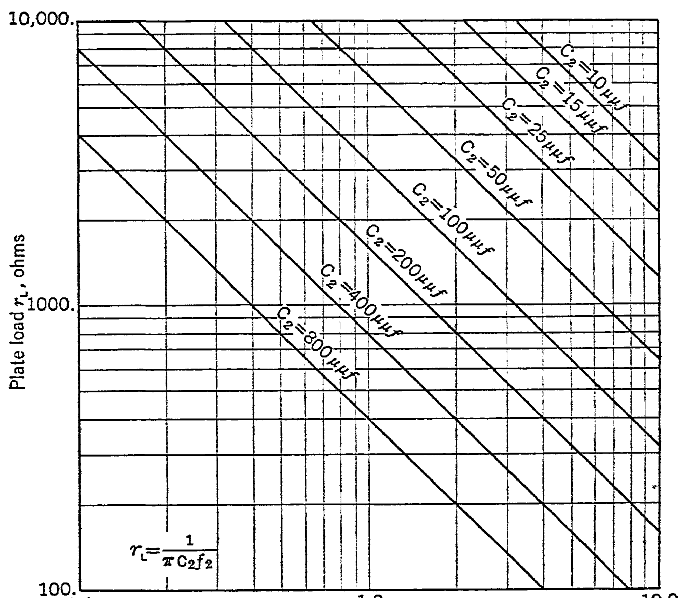


FIG. 18. Plate Load Resistance (r_L) in Terms of Frequency (f_2) and Input Capacitance (C_2) for a Constant-K Type Low-pass Coupling Network

This gives a ratio for C_d/C_1 of 0.25; referring to curve Fig. 16 M is found to be 0.78. Referring to Fig. 17 at the point M = 0.78 the value of r_L will be 78 per cent of that for a constant-K network and the value of L_2 will be 60.8 per cent of that for a constant-K network. These factors modify the value obtained above, r_L becomes 3300 ohms, and L_2

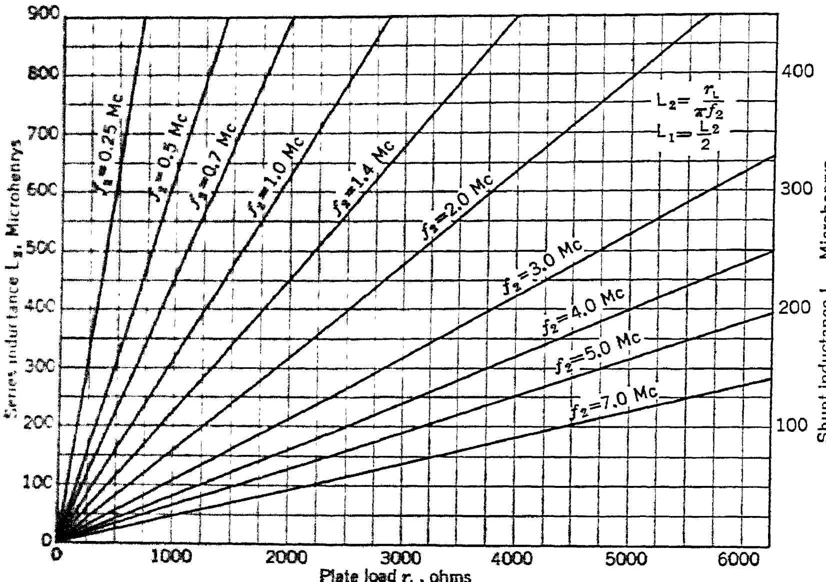


FIG. 19 Shunt and Series Inductances (L_1) and (L_2) in Terms of Plate Load Resistance (r_L) and Frequency (f_2) for a Constant-K Type Low-pass Filter-coupling Network

becomes 272 microhenrys. From the above derivation L_1 = 218 microhenrys, which is approximately that obtained from Fig. 19. The shunting resistor r_2 would have a value somewhere between 15,000 ohms and 33,000 ohms, depending upon the Q of the peaking coil (L_2). Its exact value would have to be determined by test.

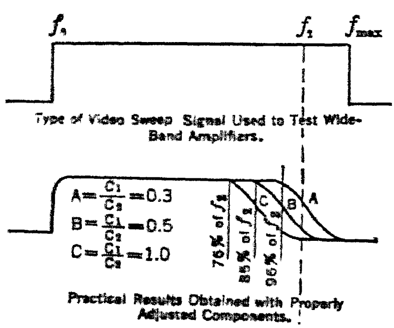


FIG. 20. Practical Results Obtained with Properly Adjusted Components (Constant-K Type)

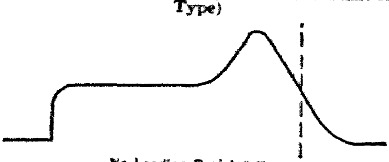


FIG. 21. Curve Shape Obtained When No Loading Resistor (r_2) Is Used

FIG. 22. Curve Shape Obtained When the Loading Resistor (r_2) Is Too Large

FIG. 23. Curve Shape Obtained When the Loading Resistor (r_2) Is Too Small

FIG. 24. Curve Shape Obtained When the Plate Load Resistor (r_L) Is Too Large

In practice the wide-band amplifier is built as designed, and then tested by means of a video sweep of constant amplitude from zero to possibly 10 megacycles. This signal is applied to the grid of the wide-band amplifier. The output of the second stage V_2 is observed across a very low plate resistance, possibly 100 ohms. Figure 20 shows the response of the wide-band amplifier of the constant-K configuration when all its components are adjusted properly. Figures 21 to 28 inclusive give the response obtained with various components that have incorrect values.

FIG. 25. Curve Shape Obtained When the Plate Load Resistor (r_L) Is Too Small

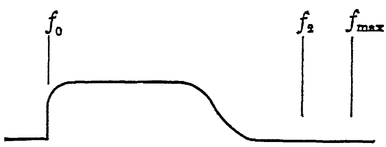


FIG. 26. Curve Shape Obtained When the Series Peaking Coil (L_2) Is Too Large

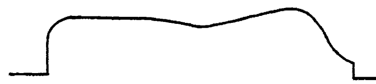


FIG. 27. Curve Shape Obtained When the Series Peaking Coil (L_2) Is Too Small

FIG. 28. Curve Shape Obtained When the Shunt Peaking Coil (L_1) Is Incorrect

Multistage wide-band amplifiers using pentodes or beam power tubes offer no particular difficulty. However, if a multistage triode amplifier is to be designed, trouble will be encountered if peaking is used in the grid circuit and plate circuit of the same tube, as it may result in a tuned-grid, tuned-plate oscillator.

FIGURE OF MERIT FOR WIDE-BAND AMPLIFIER TUBES. The capability of a vacuum tube to amplify at high frequencies depends not only upon its mutual conductance but also upon its input and output capacitances. It is these capacitances, in addition to the stray capacitance, that limits the value of plate resistance r_L that can be used. The "Miller" capacitance effect is the chief reason triodes are not satisfactory as wide-band amplifiers.

Considering these factors, eq. (29) gives an acceptable figure of merit (F.M.) for wide-band amplifiers.

$$\text{F.M.} = \frac{g_m}{C_{gk} + C_{pk} + C_{gp}(1 + A)} \tag{29}$$

Equation (29) is applicable to either triodes or pentodes; for pentodes and beam power tubes C_{gp} is so small that the last term may be ignored, hence:

$$\text{F.M.} = \frac{g_m}{C_{\text{in}} + C_{\text{out}}} \tag{30}$$

Table 1 gives a list of vacuum tubes with their corresponding figure of merit.

Table 1. List of Vacuum Tubes Applicable to Wide-band Amplifier Service, and Their Figure of Merit

Tube Type	Input Capacitance, $\mu\mu$f	Output Capacitance, $\mu\mu$f	G_m	F.M.
6AK5	4.0	2.8	5,100	750
6AG5	6.5	1.8	5,000	602
6AC7/1852	11.0	5.0	9,000	562
6AG7	13.0	7.5	11,000	536
6AU6	5.5	5.0	5,200	495
6BA6	5.5	5.0	4,400	419
6AB7/1853	8.0	5.0	5,000	385
6SH7	8.5	7.0	4,900	316
6SG7	8.5	7.0	4,700	303
6L6	10.0	12.0	6,000	273
6V6GT	9.5	7.5	4,100	241
954	3.4	3.0	1,400	219
6K6GT	5.5	6.0	2,300	200

COMPARISON OF HIGH-FREQUENCY COMPENSATION METHODS. Table 2 gives the essential design data for high-frequency compensation of wide-band amplifiers. The last three types listed as "practical results" give data obtained by measuring the constant-K network response, with different ratios of C_1/C_2. As the distributed capacitance C_d cannot be eliminated, it must be considered as producing an M-derived low-pass filter network.

Table 2. Summary of Wide-band Amplifier Formulas

Type	r_L	L_2	L_1	Relative Gain at f_2	M
Uncompensated	$\frac{1}{2\pi f_2 C_T}$	....	....	0.707	....
Shunt	$\frac{1}{2\pi f_2 C_T}$	....	$\frac{0.5 r_L}{2\pi f_2}$	1.00	....
Shunt for best phase and ampl.	$\frac{0.85}{2\pi f_2 C_T}$	....	$\frac{0.35 r_L}{2\pi f_2}$	0.85	....
Series $C_1/C_2 = 0.5$	$\frac{1.5}{2\pi f_2 C_T}$	$\frac{r_L}{2\pi f_2}$	....	1.5	....
Shunt-series $C_1/C_2 = 0.5$	$\frac{1.8}{2\pi f_2 C_T}$	$0.52 C_T r_L^2$	$0.12 C_T r_L^2$	1.8	....
Pure constant K $C_1/C_2 = 0.5$	$\frac{1}{\pi f_2 C_2}$	$r_L^2 C_2$	$0.5 L_2$	3.0	....
Constant K with C_d $C_1/C_2 = 0.5$	$\frac{M}{\pi f_2 C_2}$	$r_L^2 C_2$	$0.8 L_2$ app.	$3.0M$	$\sqrt{1 + \left(\frac{C_d}{C_1}\right)^2} - \frac{C_d}{C_1}$
		Practical Results			
Constant K with C_d $C_1/C_2 = 0.3$	$\frac{0.96M}{\pi f_2 C_2}$	$r_L^2 C_2$	$0.8 L_2$ app.	$2.88M$	$\sqrt{1 + \left(\frac{C_d}{C_1}\right)^2} - \frac{C_d}{C_1}$
Constant K with C_d $C_1/C_2 = 0.5$	$\frac{0.85M}{\pi f_2 C_2}$	$r_L^2 C_2$	$0.8 L_2$ app.	$2.55M$	$\sqrt{1 + \left(\frac{C_d}{C_1}\right)^2} - \frac{C_d}{C_1}$
Constant K with C_d $C_1/C_2 = 1.0$	$\frac{0.76M}{\pi f_2 C_2}$	$r_L^2 C^2$	$0.8 L_2$ app.	$2.28M$	$\sqrt{1 + \left(\frac{C_d}{C_1}\right)^2} - \frac{C_d}{C_1}$

Gain $= r_L G_M$

LOW-FREQUENCY RESPONSE. Low-frequency attenuation or low-frequency amplitude distortion and phase shift may be introduced in any one of four places, or a combination of the four. They are (1) cathode resistor and by-pass condenser; (2) grid condenser-resistor coupling network; (3) screen supply resistor and by-pass; (4) the internal impedance of the B power supply.

EFFECT OF A CATHODE RESISTOR AND BY-PASS. One method of obtaining a negative bias on the grid of a vacuum-tube amplifier is to include a resistor r_k in series with the cathode to ground; to prevent loss of gain, this resistor r_k is shunted by a capacitor C_k. See Fig. 29. The effect of this bias network $r_k C_k$ on the low-frequency response is caused by the fact that, the lower the frequency, the higher the capacitive reactance of C_k and the less its shunting effect on r_k. This results in cathode degeneration with an accompanying loss in gain. The gain of such an amplifier stage is given by:

$$\text{Gain } A = \frac{g_m r_L}{1 + \frac{r_L}{r_p} + Z_k\left(\frac{1}{r_p} + g_m\right)} \tag{31}$$

The cathode circuit impedance Z_k, is given by

$$Z_k = \frac{r_k/(2\pi f C_k)}{\sqrt{r_k^2 + (1/2\pi f C_k)^2}} \tag{32}$$

Equation (31) is a general formula. For a wide-band amplifier the tube is usually of the pentode type, in which case $r_p \gg r_L$, and $r_p \gg 1.0$, so r_L/r_p and $1/r_p$ may be disregarded, and eq. (31) becomes:

$$\text{Gain } A = \frac{g_m r_L}{1 + g_m Z_k} \tag{33}$$

To prevent loss of gain at the lowest frequency, Z_k must remain essentially constant. In practice this may mean hundreds of microfarads for C_k, especially if r_k is low. It is possible to compensate for the loss of gain due to cathode degeneration by a plate filter r_FC_F in the plate circuit of the amplifier stage (Fig. 29). The conditions that must be

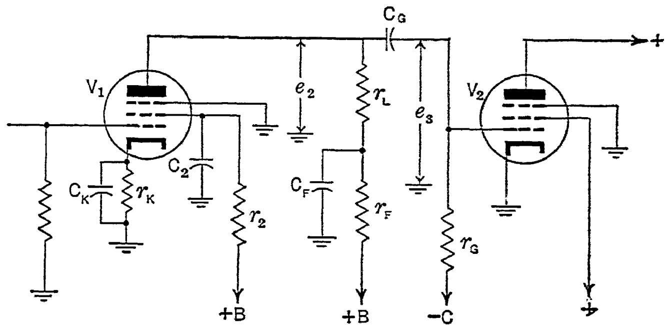


FIG. 29. Amplifier Stage Considering Low-frequency Response Only

met to compensate for the r_kC_k network are: $r_kC_k = r_FC_F$, $r_F/r_k = r_Lg_m$, and $C_k/C_F = r_Lg_m$, from which the values of r_F and C_F are obtained

$$r_F = g_m r_L r_k \tag{34}$$

and

$$C_F = \frac{C_k}{g_m r_L} \tag{35}$$

THE EFFECT OF THE GRID COUPLING CAPACITOR-RESISTOR. With a grid coupling condenser C_G and resistor r_g, Fig. 29, the voltage e_3 impressed on the grid of V_2 will decrease, as the frequency decreases, assuming e_2 to remain constant. The ratio e_3/e_2 is given by

$$\frac{e_3}{e_2} = \frac{r_G}{\sqrt{r_G^2 + (1/2\,\pi f C_G)^2}} \tag{36}$$

At the frequency that makes the capacitive reactance $1/\omega C_G$ equal to the grid resistance r_G, the response e_3/e_2 will be 70.7 per cent of that at mid-range frequency, and the phase shift will be

$$\tan^{-1}\frac{X_{C_G}}{r_G} = 45 \text{ deg} \tag{37}$$

The grid resistor r_G should be made as large as possible; its value, however, is limited by the tube manufacturer to a maximum for a given tube type. For a given low-frequency response the value of C_G may have to be so large that there is danger of ruining the high-frequency response by increased stray capacitance to ground. In practice the value of C_G would be 0.05 to 0.10 μf, and r_G would be the manufacturer's maximum value for the tube type being used. If these values do not give a response at the lowest frequency of $5f_1$ as indicated on Fig. 1, compensation is needed.

Equations (36) and (37) are plotted in the form of curves in Fig. 30. This gives an easy means of determining low-frequency response before compensation. The plate filter C_Fr_F, Fig. 29, can be so proportioned that the voltage rise across r_LC_F as the frequency decreases can just compensate for the voltage loss across C_G, thus producing flat response. Also the phase angles are such as to compensate. To achieve this compensation the time constants of the plate-filter circuit and grid circuit must be equal, that is, $r_GC_G = r_LC_F$, from which

$$C_F = \frac{r_GC_G}{r_L} \tag{38}$$

and

$$r_F = \frac{10}{2\pi f C_F} \tag{39}$$

where f is the lowest frequency to be compensated to full response. These equations are based on the fact that the pentode amplifier is a constant current device within the range of operation. The value of r_F should not be made so high that the amplifier plate voltage falls appreciably below its screen voltage, or trouble may result from too high a screen dissipation.

Usually it is preferable to make the cathode circuit such that no compensation is needed at the lowest frequency required. Then compensate the grid coupling network in the plate circuit by means of the plate filter. It must be remembered that the plate filter compensation will take care of only the loss of lows at one point. *Do not try to compensate for two losses by a single compensation;* it cannot be done.

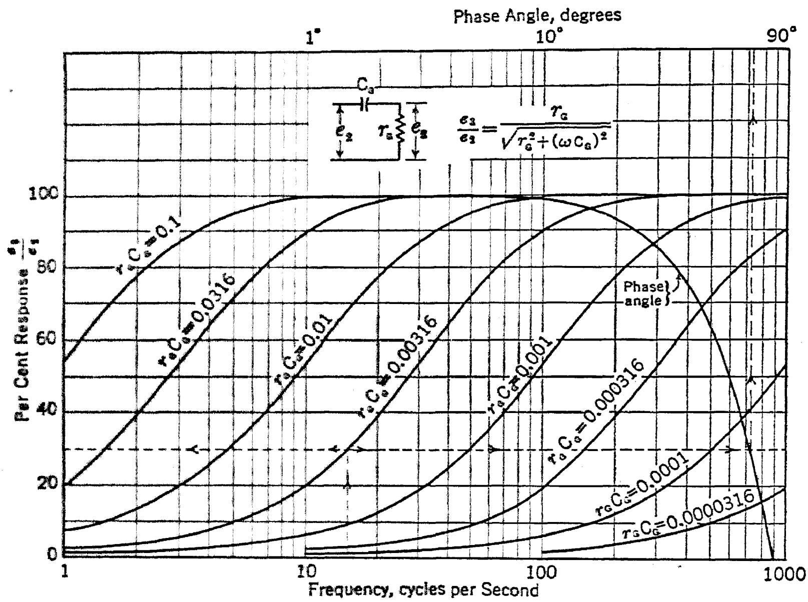


FIG. 30. Curve, Giving the Low-frequency Response Obtained, and the Low-frequency Phase Shift, with Different (r_GC_G) Grid Coupling Network

THE EFFECT OF THE SCREEN BY-PASS. The effect of r_2C_2 (Fig. 29) is similar to that which results from cathode degeneration mentioned previously. However, the screen current is only about 10 per cent of the plate current, and the screen-plate mutual conductance is only about 12 per cent of the control grid or cathode-to-plate mutual conductance, and so the effect is much smaller.

If the time constant of r_2C_2 is made at least 4 times as long as the period of the lowest frequency it is desired to pass, the effect will be negligible. The value of r_2 is determined by the voltage requirements for the particular tube being used; then the value of C_2 is given by:

$$C_2 = \frac{4}{r_2 f} \tag{40}$$

THE EFFECT OF THE INTERNAL IMPEDANCE OF THE POWER SUPPLY. The power supply internal impedance Z_b is essentially the reactance of the output filter condenser. This reactance becomes common for all amplifier stages and may result in the loss of low-frequency response or in low-frequency motor-boating, depending upon the number of stages, the response, and the gain of the system.

It has been found that the capacitive reactance x_c of the final filter condenser in the power supply system, at the lowest frequency encountered, should be no greater than 10 per cent of the effective full load resistance to prevent common coupling through the B supply. The value of the final filter capacitance is then given by:

$$C = \frac{10}{2\pi f r} \tag{41}$$

where r is the effective load resistance which is given by:

$$r = \frac{E_b}{I_b} \tag{42}$$

where E_b and I_b are the B voltage and B current of the B supply.

In general, when designed for low-frequency response, the cathode circuit is by-passed with a capacitor of sufficient size to prevent cathode degeneration at the lowest frequency encountered, or the cathode is grounded and negative bias is supplied to the grid. The screen circuit is adequately by-passed so that no loss of lows results, and the B supply impedance is made sufficiently low to prevent trouble. This leaves only the effect of the grid capacitor-resistor network which must be compensated. This compensation is done in the plate circuit of tube V_1 driving the grid of V_2, as previously explained.

6. CATHODE FOLLOWERS

The name "cathode follower" is given to an amplifier stage when the load, or the major portion of the load, is in the cathode circuit instead of in the plate circuit. Figure 31 shows such an amplifier stage.

This type of operation is called a cathode follower because the cathode tends to follow the grid in voltage as signal is applied, thus reducing the actual grid to cathode voltage

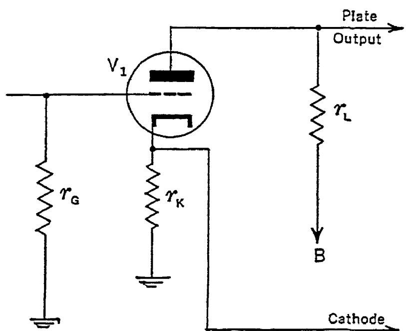


FIG. 31. Cathode-follower Stage General Case, with Resistance in Both the Plate and the Cathode Circuits

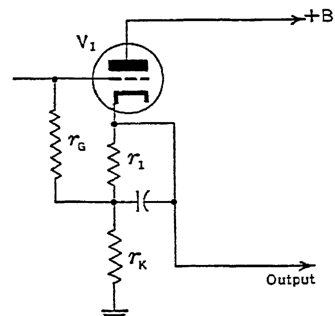


FIG. 32. Cathode-follower Arranged for Proper Bias by Returning the Grid Resistor to a Tap on the Cathode Resistor

below that of the applied signal. This type of circuit finds its widest use as an impedance-changing device. It is used extensively in connection with wide-band amplifiers to match a line where a transformer would be impractical.

Cathode followers usually use triodes, since high gain cannot be realized and since the shunt capacitance of a pentode is high, including screen-to-plate and screen by-pass capacitance as well as the usual cathode capacitance.

Cathode followers have many characteristics not found in amplifiers of other types: (1) output from the cathode circuit, (2) voltage gain to the cathode of less than 1, (3) reduction of input capacitance, (4) increase in input resistance, (5) low output impedance, (6) increase in effective plate impedance, (7) no change in polarity of output signal, and (8) use as a phase splitter with load in both cathode and plate.

When a tube is operated as a cathode follower, the circuit should be so arranged as to supply the proper negative bias for the type of tube being used. If the circuit is as shown in Fig. 31, the Ir drop across the cathode resistor r_k should be such as to provide the proper bias for class A operation. When a high value of cathode resistor is needed, the arrangement shown in Fig. 32 should be used. The Ir drop across the resistor r_1 should provide the proper negative bias for class A operation. If the cathode resistor is so low in value that sufficient negative bias is not developed, an external negative voltage must be supplied to the grid to produce proper class A operation.

In the general case (Fig. 31) the gain to the cathode is always less than 1 and is:

$$\text{Gain } A = \frac{g_m r_k}{1 + (r_L/r_p) + r_k[(1/r_p) + g_m]} \tag{43}$$

If the plate resistor is zero (Fig. 32), eq. (43) becomes

$$\text{Gain } A = \frac{g_m r_k}{1 + r_k[(1/r_p) + g_m]} \tag{44}$$

When the cathode load is not a pure resistance, r_k is replaced by Z_k, which is the absolute value of impedance at the frequency of interest.

For pentode operation $r_p \gg r_L \gg 1$, and so the gain to the cathode becomes

$$\text{Gain } A = \frac{g_m r_k}{1 + r_k g_m} \tag{45}$$

The effective input capacitance of a cathode follower is given by

$$C_{\text{eff}} = C_{gk}\left(1 - \frac{g_m r_k}{1 + r_k[(1/r_p) + g_m]}\right) + C_{gp}\left(\frac{g_m r_L}{1 + (r_L/r_p) + r_k[(1/r_p) + g_m]}\right) \tag{46}$$

When the plate resistor is zero, eq. (46) becomes:

$$C_{\text{eff}} = C_{gk}\left(1 - \frac{g_m r_k}{1 + r_k[(1/r_p) + g_m]}\right) + C_{gp} \tag{47}$$

The effective input resistance is given by:

$$r_{\text{eff}} = \frac{r_g}{1 - \dfrac{g_m r_k}{1 + (r_L/r_p) + r_k[(1/r_p) + g_m]}} \tag{48}$$

The effective output impedance r_0 of a cathode follower, for the general case, is given by:

$$r_0 = \frac{(r_L/r_p) + 1}{(1/r_p) + g_m} \tag{49}$$

Equation (49) is for the tube alone and does not include the effect of the cathode resistor r_k which is in shunt with r_0. The resulting impedance Z_1 is given by:

$$Z_1 = \frac{r_0 r_k}{r_0 + r_k} \tag{50}$$

When the plate resistor r_L is zero, eq. (49) becomes:

$$r_0 = \frac{1}{(1/r_p) + g_m} \tag{51}$$

In a pentode $r_p \gg 1$, and so $r_0 = 1/g_m$.

In some cases it may be more desirable to have Z_1 the effective output impedance of a cathode follower in a single equation rather than first to calculate r_0 and then Z_1. The general case for Z_1 directly is

$$Z_1 = \frac{r_L + r_p}{1 + r_p[g_m + (1/r_k)] + (r_L/r_k)} \tag{52}$$

When r_L is zero, eq. (52) becomes:

$$Z_1 = \frac{1}{(1/r_k) + (1/r_p) + g_m} \tag{53}$$

LINE MATCHING. It is often necessary to couple into a line, matching its characteristic impedance. For audio frequencies this can be done with a matching output transformer, but for wide-band amplifier work such a transformer is not readily available. A cathode follower can be used for such service.

When the characteristic impedance Z of the line to be matched is less than r_0 for the tube to be used the value of the cathode resistor r_k needed to match can be found from eq. (54); the circuit is shown in Fig. 33.

$$r_k = \frac{Z r_p}{r_p - Z(1 + g_m r_p)} \tag{54}$$

When the characteristic impedance Z of the line is higher than the cathode impedance r_0 of the tube, a series resistor is needed (Fig. 34), its value being the difference between Z and r_0 from eq. (51).

In calculating the gain of a cathode follower feeding a line as in Fig. 33, the resulting impedance on the tube cathode is the cathode resistor r_k and the line impedance Z in parallel. This resultant impedance Z_k is used in eq. (44) to obtain the gain of the stage.

In the second case, where the line characteristic impedance Z is greater than r_0, as in Fig. 34, the gain to the line is

$$\text{Gain to line} = \left(\frac{Z}{r_s + Z}\right)\left(\frac{g_m(r_s + Z)}{1 + (r_s + Z)[g_m + (1/r_p)]}\right) \tag{55}$$

When matching a line with a characteristic impedance Z higher than the tube cathode impedance r_0 as in Fig. 34, there must be a d-c circuit through the line capable of carrying the tube plate current; otherwise the cathode follower will not operate. If r_k is too high, however, the operation may be improved by terminating the line of Fig. 33 for direct current as well as for alternating current.

When an amplifier operates with an unbypassed cathode resistor (a type of cathode follower), the effective plate resistance r_p' is increased.

$$r_p' = r_p(1 + g_m r_k) \tag{56}$$

When a cathode follower is operated with a resistor in the plate circuit equal to the resistor in the cathode circuit, it is termed a "phase splitter." That is, the voltage developed in the plate circuit will be equal to the voltage developed in the cathode circuit.

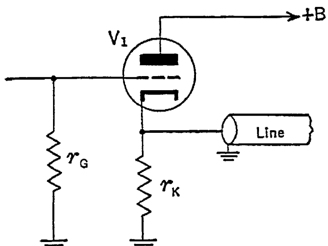


FIG. 33. Cathode-follower Circuit Used to Match a Transmission Line, When the Characteristic Impedance (Z) of the Line Is Lower Than the Cathode Impedance (r_0) of the Tube (V_1)

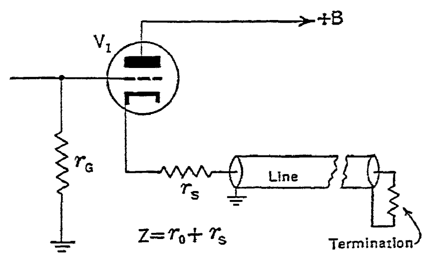


FIG. 34. Cathode-follower Circuit Used to Match a Transmission Line, When the Characteristic Impedance (Z) of the Line Is Higher than the Cathode Impedance (r_0) of the Tube (V_1)

However, the polarity of the voltage in the plate circuit will be the inverse of the voltage in the cathode circuit. This type of circuit may be used to obtain pushpull operation from a single amplifier. The gain to the plate will be given by eq. (31), and the gain to the cathode will be given by eq. (43).

7. GROUNDED-GRID AMPLIFIER

A grounded-grid amplifier is usually a triode which has its grid grounded and the input connected to the cathode. The output circuit is in the plate in the usual manner. Figure 35 shows such a circuit. A triode connected in this manner has some of the characteristics of a screen-grid tube, as the grid acts as a shield between the input and output circuits. It also has some of the characteristics of a cathode follower, as the cathode-input impedance is equivalent to that of a cathode follower with a plate load. Another feature of the cathode-input amplifier is that there is no voltage inversion between the input signal and the output signal as with conventional grid input.

The input impedance Z_1 is given by:

$$Z_1 = \frac{r_L + r_p}{1 + r_p[g_m + (1/r_k)] + (r_L/r_k)} \tag{52}$$

The input impedance Z_1 of a cathode-input amplifier may be adjusted to match a line the same as a cathode follower. Solving eq. (52) for r_k gives

$$r_k = \frac{1 + (r_L/r_p)}{1/2\,[1 + (r_L/r_p)] - [g_m + (1/r_p)]} \tag{57}$$

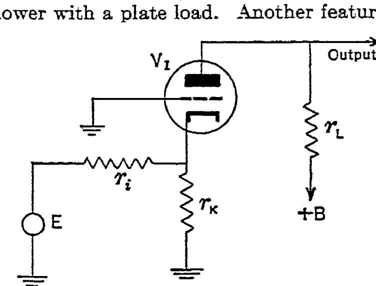


FIG. 35. Circuit of the Grounded-grid Amplifier

The effective impedance Z_k in the cathode circuit must be determined before it is possible to obtain the gain to the plate and the effective plate resistance r_p'. Referring to Fig. 35, r_i represents the internal impedance of the signal input source, and r_k is the cathode resistor.

$$Z_k = \frac{r_i r_k}{r_i + r_k} \tag{58}$$

Where a transmission line is being matched, r_i would be replaced by Z, the characteristic impedance of the line.

The gain of the cathode-input amplifier is given by eq. (31) using Z_k as obtained from eq. (57).

$$\text{Gain } A = \frac{g_m r_L}{1 + (r_L/r_p) + Z_k[(1/r_p) + g_m]} \tag{31}$$

It can be seen from eq. (31) that to obtain high gain with a cathode-input amplifier the cathode impedance Z_k should be kept low and r_L should be made as high as possible. Making r_L high also increases the input impedance as shown by eq. (52).

The effective plate impedance r_p' of a cathode-input amplifier is given by eq. (56), substituting Z_k as obtained from eq. (58)

$$r_p' = r_p(1 + g_m Z_k) \tag{56a}$$

which gives an increase in plate resistance over that of the tube operated in the conventional manner.

8. IN-PHASE AMPLIFIERS

The name in-phase amplifier applies to an amplifier that has the same polarity of signal in the output as that applied to the input. In sine-wave operation this is of little importance, but for pulse amplifiers, or television amplifiers, the polarity of the signal is important. These types of signals are not symmetrical about an a-c axis and must be treated accordingly.

There are four general types of in-phase amplifiers:

1. Cathode follower.
2. Cathode-input amplifier.
3. Combined cathode follower, cathode-input amplifier (cathode-coupled amplifier).
4. Suppressor input, screen output amplifier.

The cathode follower and cathode-input-type amplifiers have been covered in the preceding sections and will not be discussed further.

The third type of in-phase amplifier is a combination of cathode follower and cathode-input amplifier, termed cathode-coupled amplifier. This type of amplifier is best made by using a dual triode with a common cathode, such as a 6J6 or a 6SN7 tube. The circuit is shown in Fig. 36. In an amplifier of this type the polarity of the signal is unchanged at any point through the stage.

The value of r_k should be such that the negative bias developed by the combined plate currents of V_1 and V_2 is proper for class A operation of the tubes V_1 and V_2. In operation the cathode-coupled amplifier exhibits characteristics in some respects similar to those of a screen-grid tube. It can be used with tuned input and tuned output circuits without danger of oscillation, such as would occur if a triode in conventional circuits were used.

FIG. 36. The In-phase Amplifier of the Cathode-coupled Type

The gain of (V_1) the cathode follower can be calculated by using eq. (31), the value of Z_1 obtained from eq. (52) being substituted for Z_k. These two equations can be combined to give the gain to the cathode of V_1; assuming that both tubes are the same, then

$$\text{Gain to cathode of } V_1 = \frac{g_m(r_L + r_p)}{2(1 + r_p g_m) + (r_L + r_p)/r_k + r_L[g_m + (1/r_p)]} \tag{59}$$

The cathode follower is working into the cathode of the second tube which exhibits a low cathode-input impedance, as does any cathode-input amplifier. This cathode impedance must be considered in the gain equation. This is done by using Z_1 of the cathode-input tube as the cathode load for the cathode follower. This is a lower value than that of the cathode resistor.

In obtaining the gain of the cathode-input section V_2, the cathode-output impedance Z_1 for the cathode follower V_1 must be used for the cathode load of the cathode-input amplifier V_2. The value of Z_1 from eq. (53) must be substituted for Z_k in eq. (31). Combining these two equations, the gain of the cathode input section is given by:

$$\text{Gain from cathode to plate of } V_2 = \frac{g_m r_L[1 + (r_p/r_k) + r_p g_m]}{2(1 + r_p g_m) + (r_L + r_p)/r_k + r_L[g_m + (1/r_p)]} \tag{60}$$

The overall gain of the cathode-coupled amplifier is given by the product of the gain to the cathode of V_1 times the gain from the cathode to the plate of V_2. This can be written in a single equation as

Overall gain of cathode-coupled stage

$$= \frac{g_m^2 r_L(r_L + r_p)[1 + (r_p/r_k) + r_p g_m]}{\{2(1 + r_p g_m) + [(r_L + r_p)/r_k] + r_L[(1/r_p) + g_m]\}^2} \tag{61}$$

SUPPRESSOR INPUT, SCREEN OUTPUT AMPLIFIER. The circuit for this amplifier is given in Fig. 37. The circuit requirements are as follows: the tube should be of a type that normally operates with the screen at a lower voltage than the plate, and also it must have a suppressor grid structure that has an effective mutual conductance to the plate. The 6AS6 is a satisfactory tube for this type of operation, as the suppressor-grid-to-plate has a mutual conductance of about 1000 micromhos, and the suppressor grid-to-screen-grid has a mutual conductance g_{ms} of about 850 micromhos. The screen impedance Z_s is of the order of 10,000 ohms.

By the use of a 6AS6 for an in-phase amplifier, the values shown in Fig. 37 give satisfactory results. Care must be taken when operating a tube in this manner not to exceed the screen dissipation.

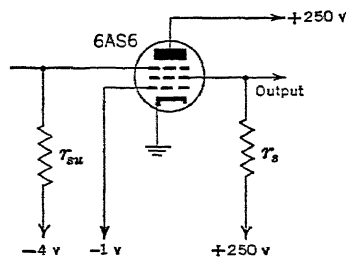


FIG. 37. The In-phase Amplifier of the Suppressor Input Screen Output Type

The gain A to the screen is given by eq. (3) using r_s in place of Z and the screen impedance Z_s in place of r_p; the equation then becomes

$$\text{Gain } A = \frac{g_{ms} r_s Z_s}{r_s + Z_s} \tag{62}$$

where g_{ms} is the mutual conductance of the suppressor to the screen.

9. NEGATIVE-FEEDBACK AMPLIFIERS

In the negative-feedback amplifier a voltage obtained from the amplifier output is fed back to the input in such a way as to oppose the applied signal.

There are two general types of negative feedback. The first is negative voltage feedback which occurs when a fraction β of the output proportional to the voltage across the output load is fed back to the input. The second is negative current feedback, which occurs when the voltage fed back is proportional to the current through the output load. The major difference in the results produced by voltage and current feedback are that negative voltage feedback results in a reduction of the effective internal resistance of the amplifier, whereas negative current feedback produces an effective increase in the internal resistance of the amplifier.

Owing to the effect of reactance in the circuit, the voltage which is fed back may not be wholly out of phase with the input voltage. This phase shift is likely to occur at extremely low or extremely high frequencies. In a single stage it cannot exceed 90°, which results in no feedback and a corresponding increase in gain. When more than one stage is included in the feedback amplifier, the phase shift may exceed 90°, which results in regeneration and perhaps oscillation. The method used to combat this phase shift and resulting oscillation is to make one stage with a narrower band width than the others. This should result in a loss of gain through the narrow stage to a value at which oscillation cannot occur, by the time the phase shift has exceeded 90°.

The effects of negative voltage feedback are (1) reduction in gain, (2) reduction in distortion, (3) reduction in noise, (4) improvement in the fidelity with frequency, (5) greater consistency of characteristics with changes in applied voltages, and (6) reduction of the effective internal resistance of the final amplifier stage.

The gain or amplification in the presence of voltage feedback is given by the relation

$$\text{Gain with feedback } A' = \frac{A}{1 - A\beta} \tag{63}$$

However, for negative feedback β is negative and the relations become

$$\text{Gain with negative feedback } A' = \frac{A}{1 + A\beta} \tag{64}$$

which is a reduction in gain from A, the amplification without feedback.

When the value of A is large in comparison to 1, the gain becomes practically independent of the amplifier characteristics, becoming approximately $A' = 1/\beta$.

Negative voltage feedback reduces the non-linear harmonic distortion produced in the amplifier for a given output voltage according to the relation

$$(D') \text{ Distortion with negative feedback} = \frac{D}{1 + A\beta} \qquad (65)$$

where D is the distortion with no feedback. This assumes that no distortion is produced when reamplifying the distortion voltages fed back, which is quite accurate if the distortion with no feedback is not large.

Feedback will reduce the distortion up to a certain point, but feedback cannot increase the power-output capabilities of a given amplifier. The distortion will be low over a portion of the output range but then will increase faster than for an amplifier with no feedback (Fig. 38).

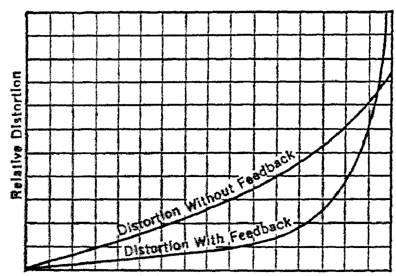


FIG. 38. The Effect of Negative Voltage Feedback on Distortion

Negative voltage feedback will improve the signal-to-noise ratio, if the source of noise is in the amplifier and not fed in as part of the input signal. For this case, assuming equal output voltages,

$$\frac{\text{Signal to noise with feedback}}{\text{Signal to noise without feedback}} = \frac{A_1}{A_2(1 + A\beta)} \qquad (66)$$

where A_1 is the amplification from the point of introduction of the noise to the output with negative feedback, and A_2 the amplification from the point of noise introduction to the output without feedback.

With negative feedback, as the response starts to drop with either high or low frequency, the feedback also decreases, opposing the change and resulting in a flatter frequency response. Thus

$$\frac{\text{Gain at } f_1(A_1)}{\text{Gain at } f_2(A_2)} = \frac{A_1(1 + A_2\beta)}{A_2(1 + A_1\beta)} \qquad (67)$$

Negative voltage feedback causes an apparent reduction in the plate resistance of the final amplifier tube in a feedback amplifier. The actual plate resistance r_p does not change, so any impedance matching (as with an output transformer) should be done on the basis of no feedback. Then, with feedback, the response and damping effect as on a loudspeaker resonance is the same as though the plate resistance r_p were lowered to r_p'. The effective plate resistance r_p' is the same whether the feedback is over a single stage or several stages, provided the gain reduction is the same.

The effective plate resistance r_p' is given by

$$r_p' = \frac{1}{g_m\beta + (1/r_p)} = \frac{r_p}{1 + \beta\mu} \qquad (68)$$

Figure 39 shows some typical circuits with negative voltage feedback applied to a single stage of amplification.

Current feedback has some characteristics that make it less desirable than voltage feedback for use with the final amplifier stage. The effects of negative current feedback are: (1) reduction of gain, (2) increase of the effective internal plate resistance r_p' of the final amplifier stage, (3) increase of input resistance, and (4) reduction of input capacitance.

The gain of such an amplifier, A', taking feedback into consideration, is given by

$$A' = \frac{A}{1 - A\alpha} \tag{69}$$

but, since α has a negative sign for negative feedback, eq. (69) in reality should be

$$A' = \frac{A}{1 + A\alpha} \tag{70}$$

where α is the ratio of the feedback resistor r_n to the load resistance r_L, and A is the amplifier gain without feedback.

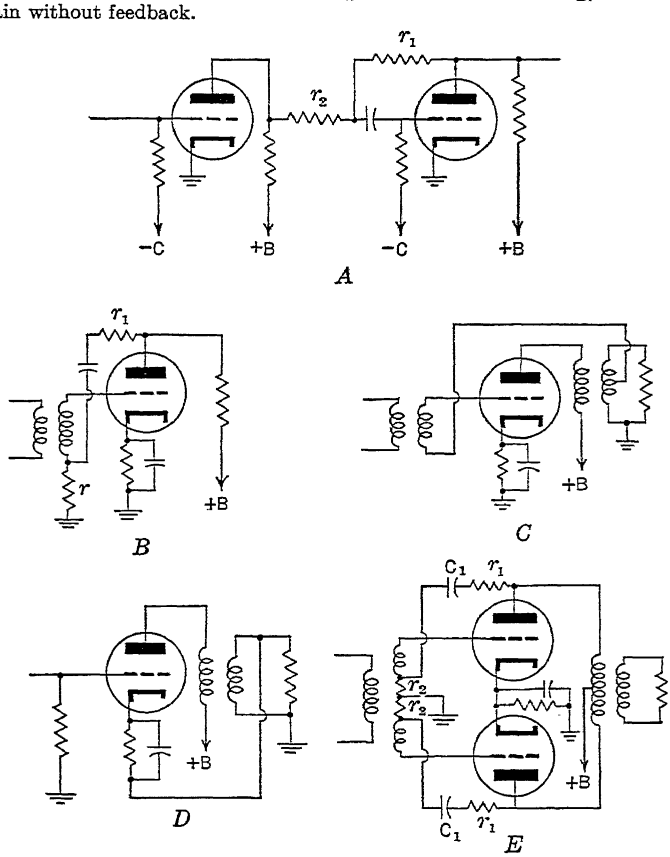


FIG 39. Typical Negative Voltage Feedback Circuits Applied to a Single Stage of Amplification

The effective plate resistance r_p' is given by the expression

$$r_p' = r_p(1 + g_m r_n) \tag{71}$$

where g_m is the mutual conductance of the amplifier output stage.

The increase of input resistance is similar to that obtained with a cathode follower. Equation (48) will give the effective input resistance; r_k should be replaced with r_n.

The input capacitance is given by eq. (46) by replacing r_k with r_n. Equations (46) and (47) apply to current feedback over a single stage.

Figure 40 shows some typical circuits with negative current feedback.

THE ONE-SHOT AMPLIFIER. The one-shot amplifier is a form of multivibrator with one of the tubes biased beyond cutoff. The circuit is shown in Fig. 41. Under steady-state conditions tube V_1 is cut off and tube V_2 draws plate current. The circuit remains in this condition until a positive trigger voltage of sufficient amplitude to cause V_1 to draw plate current is impressed upon the grid of V_1. This starts a regenerative action through V_2 and the grid of V_1 is driven positive (as indicated in Fig. 42). This causes grid current to flow, thus charging the coupling capacitances C_1 and C_2, resulting in a large negative potential on the grid of V_1. As soon as the current in the plate circuit of V_1 stops increasing, the voltage on the grid of V_2 starts to rise; this is also regenerative, and V_2 resumes its steady-state plate current, but V_1 is left with its grid highly negative. This negative voltage discharges through r_1 exponentially back to the applied bias voltage E_2, Fig. 42.

Owing to the regenerative action the output of this type of amplifier is a sharp pulse of short duration. The plate of V_1 gives a negative pulse, and the output from V_2 is a positive pulse.

The voltages acting on the grid of V_1 are shown in Fig. 42. The trigger pulses should be limited in amplitude so that, at the desired time after firing, the amplifier is again ready to fire. With the trigger pulse limited to amplitude P it cannot fire the one-shot amplifier earlier than the prescribed time. Assuming that C_1 and C_2 are small enough to permit them to change to the peak positive voltage applied, then the operating conditions can be calculated.

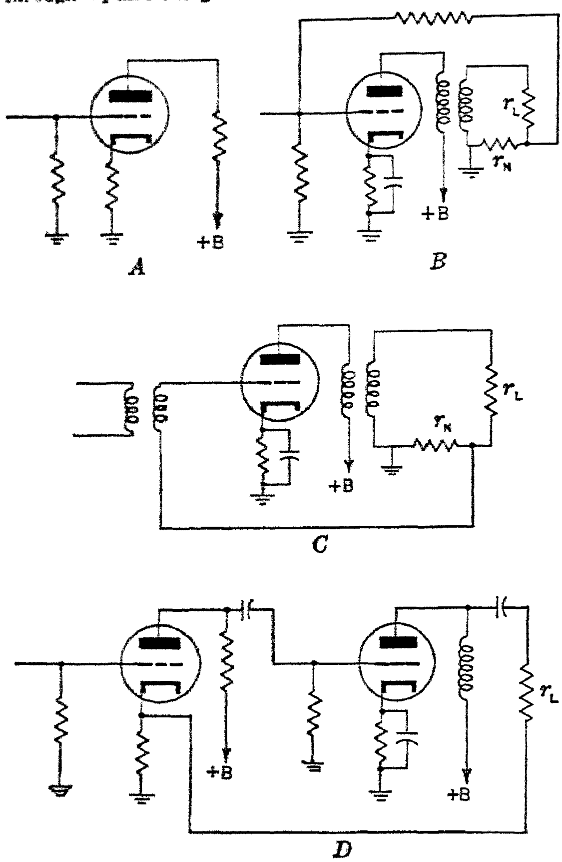


FIG. 40. Typical Negative Current Feedback Circuits

The Ir drop through the plate load resistor r_L of V_2 must be appreciably greater than the cutoff voltage required on V_1.

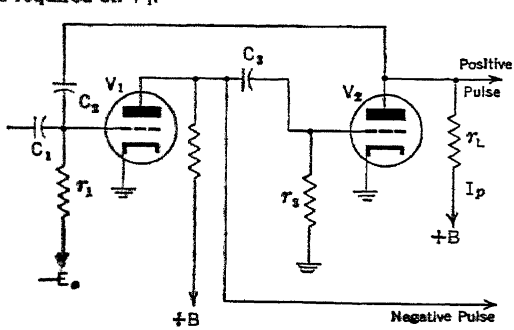


FIG. 41. Circuit Diagram of a One-shot Amplifier

Referring to Fig. 42 for voltages and symbols, assume the trigger pulse amplitude P to be equal to the cutoff bias E_1 of V_1, and the applied bias to be equal to $E_1 + P/2$.

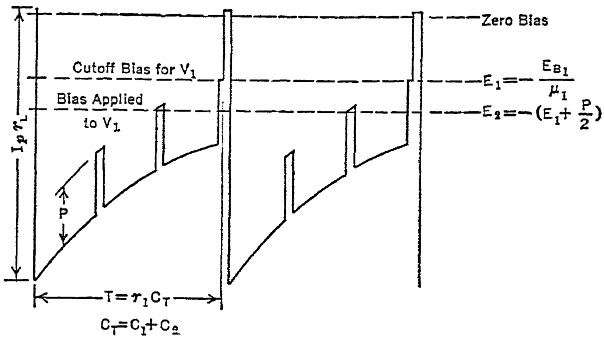


FIG 42. The Voltage on the Grid of Tube (V_1) of Fig. 41 during Operation

The time T that must elapse between firing of the amplifier and the second firing by a pulse of amplitude P is $T = r_1(C_1 + C_2)$. The applied bias $-E_c$ to V_1 of Fig. 41 is given by:

$$E_c = -\left(\frac{E_{B1}}{\mu_1} + \frac{P}{2}\right) \tag{72}$$

and E_{B1} is given by:

$$E_{B1} = \mu_1(r_L I_p - 1.85P) \tag{73}$$

where μ_1 is the amplification constant of V_1, and $r_L I_p$, the voltage drop across the plate load r_L of V_2, is assumed to be the maximum negative voltage on the grid of V_1 immediately after firing.

10. PULSE AMPLIFIER

With the advent of radar and television, pulse amplifiers became necessary. The pulse amplifier is an adaptation of the wide-band amplifier. The bandwidth necessary is dependent upon the pulse, the high-frequency response is governed by the rate of rise and decay of the pulse, and the low-frequency response is determined by the duration of the pulse. (See Section 9.)

The equivalent frequency f_2 of the pulse is considered to be equal to that of a sine wave that rises from zero voltage to peak voltage in the same time as that of the pulse. This results in a frequency f_2 that must be passed by the pulse amplifier, given by

$$f_2 = \frac{1}{4T} \tag{74}$$

where T is the rise time or decay time of the pulse, whichever is shorter (see Fig. 43).

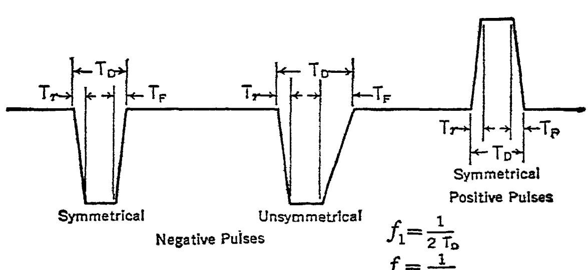


FIG. 43. Negative and Positive Pulses, Showing the Pulse Rise Time (T_r), the Pulse Decay Time (T_F), and the Pulse Duration (T_D)

Determine the frequency f_2 represented by the rise time, then refer to the section on wide-band amplifiers and determine the design of an amplifier having the required high-

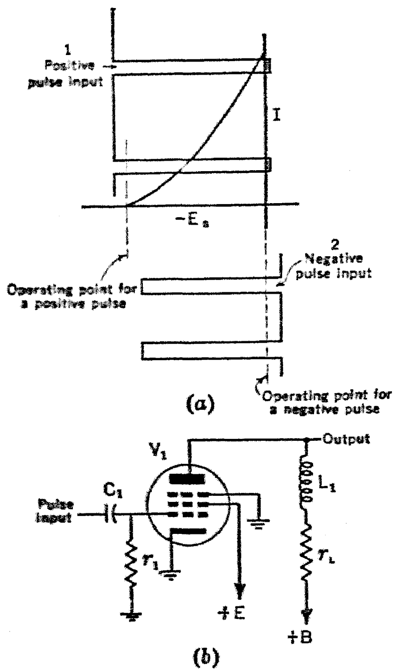


FIG. 44. Grid Bias, Plate Current Curve, Showing the Operating Point for a Positive Pulse, the Operating Point for a Negative Pulse, and a Circuit That Will Automatically Set the Bias at the Proper Voltage for Correct Operation Regardless of the Pulse Polarity or Amplitude

frequency response. The low-frequency response at $5f_1$ needed is obtained from the pulse duration T_D.

$$f_1 = \frac{1}{2T_D} \tag{75}$$

Again refer to the section on wide-band amplifiers to determine the constants needed to produce the required response.

The consideration for the initial bias point for the vacuum-tube grid is somewhat different in pulse amplifiers from what it is in conventional amplifiers. For a positive pulse a high bias is needed. This bias can be, and usually is, supplied by grid current (Fig. 44, A_1).

For a negative pulse a low bias is needed and the tube draws a heavy current except in the presence of the pulse which drives the tube grid toward cutoff. This is shown in Fig. 44, A_2.

The circuit shown in Fig. 44, B, is adequate for either a positive or a negative pulse amplifier. The time constant of the grid circuit r_1C_1 should be several times the pulse repetition rate to maintain bias between pulses.

With positive pulses the grid is biased by means of grid-leak bias to a value represented by the a-c axis of the positive pulse. This is self-adjusting and requires no controls. When one is operating with negative pulses, the bias will be essentially zero, depending upon the pulse a-c axis, and will operate equally as well as for a positive pulse. A circuit of this type is self-adjusting and assumes an operating point so as to provide efficient operation regardless of pulse amplitude or polarity.

INTERMEDIATE-FREQUENCY (I-F) AMPLIFIERS

By Charles J. Hirsch

Because it is difficult to amplify and separate signals at high radio frequencies, some receivers (known as superheterodynes) *convert the signal frequency to a fixed intermediate frequency*. The signal then is amplified and selected at the new frequency by means of an *i-f amplifier*. Such receivers differ from tuned-radio-frequency (t-r-f) receivers, which amplify by means of circuits tuned to the high carrier frequency.

The intermediate frequency is usually lower than the radio frequency and higher than the frequency of utilization (audio or video frequency).

The intermediate-frequency amplifier has the function of amplifying the signals within a specified i-f band and of rejecting all others. It is the most important factor in the determination of sensitivity, selectivity, and fidelity of superheterodyne receivers. Since these characteristics of the complete receiver are, in the main part, the characteristics of the i-f amplifier, it is easy to make them constant over the tuning range.

11. FACTORS AFFECTING THE CHOICE OF INTERMEDIATE FREQUENCY

The choice of intermediate frequency requires a careful study of the following factors: (1) overall gain, (2) selectivity, (3) image rejection, (4) tuning range, (5) tweets (whistles caused by harmonics of the i-f which are generated by the second detector and reimpressed on the r-f circuits to beat with the signal frequency), (6) i-f rejection, (7) strong stations separated by the intermediate frequencies, (8) cost, i.e., number of tuned circuits and their components.

Low intermediate frequencies have the advantages of (*a*) high stage gain because a higher impedance can be presented to the output of the amplifier tube; (*b*) narrow band width, i.e., better selectivity because a given frequency separation is a greater fraction of a low intermediate frequency than of a high intermediate frequency (see Universal selectivity curve, Section 6); and (*c*) greater freedom from tweets (when the i-f amplifier is preceded by a high degree of r-f selectivity) because only higher and therefore weaker (but more numerous) harmonics of the i-f occur at the signal frequency and beat with the signal to produce tweets.

High intermediate frequencies have the advantages of (*a*) *higher image rejection* by (1) increasing the separation (twice intermediate frequency) between the desired signal frequency and the image frequency, and (2) reducing or even eliminating that part of the tuning range within which signals can produce images; (*b*) reduction in the number of "tweets" because fewer harmonics of the intermediate frequency lie within the tuning range; (*c*) greater freedom from "birdies" (whistles produced by combinations of r-f signals) because combinations of r-f signals, separated in frequencies by the intermediate or subharmonics of the intermediate frequency, will not be impressed on the converter to produce intermediate frequency, or beat with the local oscillator to produce intermediate frequency; (*d*) greater freedom of interaction (pulling) between the local oscillator and the antenna circuit because of greater frequency separation.

High image rejection and freedom from "birdies" require costly r-f selectivity. Therefore, a high intermediate frequency is economical because it reduces the requirements for r-f selectivity.

The i-f amplifier frequency must not be too close to the tuning band as the receiver will then become unstable.

Table 1 presents a comparison of the receiver characteristics associated with two intermediate frequencies for the broadcast band; Table 2 gives some idea of the intermediate frequencies commonly associated with specific radio frequencies.

Table 1. Comparison of Two Radio Receivers Having (*a*) an Intermediate Frequency of 175 kc, (*b*) an Intermediate Frequency of 455 kc

Tuning Range 550–1720 kc	i-f 175 kc	i-f 455 kc
1. Frequency separation between desired station and image	2 × 175 = 350 kc	2 × 455 = 910 kc
2. Frequency range in which stations within the tuning range can cause images		
From	550 + 2 × 175 = 900 kc	550 + 2 × 455 = 1460 kc
To	1720 kc	1720 kc
(Note: The ability of stations outside the tuning range to produce images must not be overlooked.)		
3. Frequency range of stations which may be interfered with by images produced by stations in the tuning range. (See above.)		
From	550 kc	550 kc
To	1720 − 2 × 175 = 1370 kc	1720 − 2 × 455 = 810 kc
4. Harmonics of the intermediate frequency occurring in the tuning range	4th, 5th, 6th, 7th, 8th, 9th	2nd, 3rd
5. Separation of stations capable of beating with each other in the first detector to produce intermediate frequency	175 kc	455 kc

Table 2. Examples of Usual Intermediate Frequencies

Tuning Range, Mc	Intermediate Frequency, Mc
(*a*) 0.150–0.275	0.130
(*b*) 0.150–1.720	0.455–0.465 with some European receivers at 0.260–0.360 to gain more selectivity
(*c*) 0.540–23.0	0.455 Some receivers use 0.455 Mc for the whole frequency coverage. Others switch to an i-f of 2 Mc when receiving signals above 9 Mc.
(*d*) 40–50 Mc F-m	4–5
(*e*) 88–108 Mc F-m	10.7
(*f*) 54–88 and 174–216 Telev.	20–30 (21.75 sound—26.25 picture)
(*g*) 200 Mc pulse, communication	11.7–15–30
(*h*) 1000 Mc and up	30–60 Mc

12. NARROW- AND MEDIUM-BANDWIDTH I-F AMPLIFIERS

I-f amplifiers can be classified into (*a*) narrow (10 kc) and medium (< 200 kc) bandwidth, and (*b*) wide-bandwidth (> 200 kc) amplifiers.

Narrow- and medium-bandwidth i-f amplifiers, which are used in most receivers receiving audio frequency a-m or f-m intelligence, usually consist of individual stages using pairs of coupled-tuned circuits. A rectangularly shaped selectivity curve (one having a flat top and steep sides) usually is desirable.

I-F AMPLIFIERS FOR A-M BROADCAST RECEIVERS. These usually consist of cascaded, critically coupled, double-tuned stages at a frequency between 455 and 465 kc.

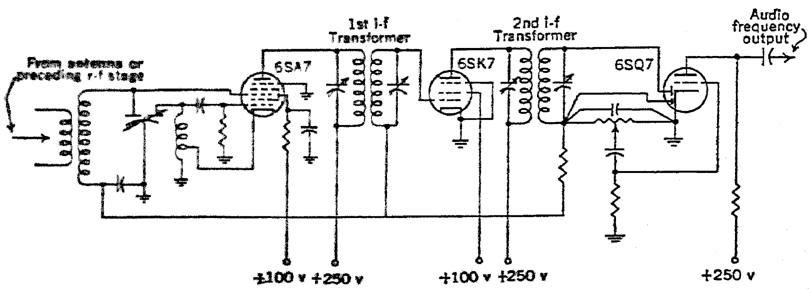


FIG. 1. Typical I-f Amplifier for Inexpensive 550–1720 kc Receiver

The design is affected by the need for (1) gain, (2) selectivity, (3) fidelity, (4) stability, and (5) economy.

Gain and Selectivity. The i-f amplifier is the major source of gain and selectivity in a radio receiver. The i-f gain (from radio frequency on the converter grid to the second detector) depends on the performance required. It will lie between a minimum of 1400 for single-band (usually 550–1750 kc) sets having a high gain r-f stage and large r-f pick-up up to a maximum of 50,000 for short-wave sets with small r-f pick-up and using low-noise converter tubes. For single-band sets it is seldom necessary to use more than one i-f stage to get the required gain and sensitivity. In general in such cases (two i-f transformers, see Fig. 1) the stage gain must be high. This requires high-impedance transformers which in turn implies high L/C ratio and high Q circuits to gain the required selectivity (see eqs.

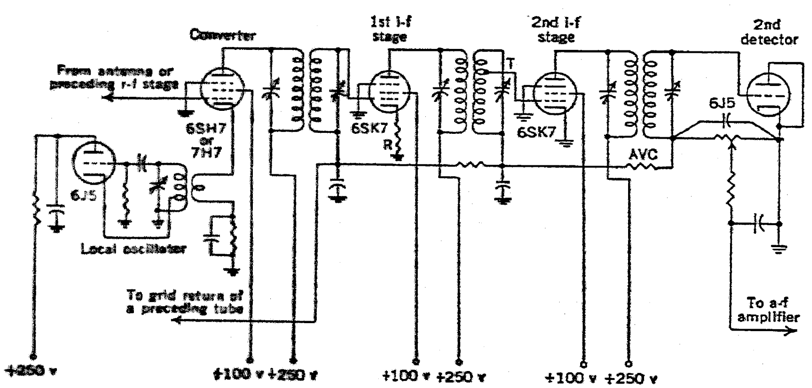


FIG. 2. Two-stage I-f Amplifier Using Six Tuned Circuits

(1)–(18) of this article). **The Q is made high** by using (*a*) litz wire, (*b*) iron-core coils, (*c*) sectionally wound (pie wound) coils, and (*d*) large shield cans.

When higher selectivity and gain are required, and in general for multiband sets, two i-f stages are often used (five to six tuned circuits, see Fig. 2). The gain and selectivity of such sets are given by eqs. (16)–(18) below. Values for typical sets are shown in Table 3.

Table 3. Average Stage Gains and Second Detector Sensitivities for Different Types of Broadcast Receivers Produced between 1934 and 1946

Type of Set \ Gain	Conversion at 600 kc μ	1st i-f μ	2nd i-f μ	Overall μ	2nd Detector Sensitivity Will Produce i-f Volts 30% Mod.	a-f Watts in Voice Coil
Ac-dc (no r-f stage)	30		94	2,800	0.3	0.05
Ac-dc (r-f gain of 6)	23		61	1,400	0.3	0.05
Battery (no r-f stage)	43		88	2,800	0.5	0.05
(no r-f stage)	32	35	44	50,000	0.5	0.05
(r-f gain of 10)	39		56	2,200	0.5	0.05
Ac (no r-f stage)	40		100	4,000	0.6–0.9	0.50
(no r-f stage)	26	10	56	14,300	0.6–0.9	0.50
(r-f gain of 12)	14	37	41	21,000	0.6–0.9	0.50
Auto (no r-f)	44		120	5,300	0.7–1.1	0.50
(r-f gain of 40)	39		90	3,500	0.7–1.1	0.50

If two i-f stages are used, the overall gain can be held down to reasonable values, stability can be improved, and the cost can be reduced by using (*a*) solid wire coils instead of litz as the five or six tuned circuits will supply adequate selectivity even with the lower Q of the solid wire coils (*b*) lower L/C ratio, (*c*) output voltage obtained from a tap on the secondary (but not in the stage feeding the diode, as a condenser across the diode load is necessary to present a low reactance to i-f harmonics generated by the diode), (*d*) unbypassed cathode resistor, (*e*) increased bias (however, this decreases the effect of avc on this tube).

Gain and Selectivity of Last Stage. The gain of the stage feeding the avc diode must be high enough so that its own grid will not overload owing to inadequate gain control of the preceding stage. In other words, it must, without overloading, supply enough power to the diode so that the diode may supply adequate control voltage to the preceding stages for all signal amplitudes to be expected. For remote cutoff tubes 30 to 40 volts of avc may well be needed, which represents from 60 to 85 per cent of the peak carrier voltage impressed on the diode.

The ratio of a-c to d-c diode load impedance should be as near unity as possible to prevent amplitude distortion on high percentage modulation signals.

Variable Selectivity. (See reference 4.) Extreme selectivity usually causes extreme cutting of the side bands with loss of fidelity. To overcome this, some receivers use variable selectivity. This usually is obtained by coupled circuits which are critically coupled (or undercoupled) when selectivity is required, but which are overcoupled when selectivity can be sacrificed. As the coupling increases, the selectivity of each coupled pair assumes the well-known two-peak form. The frequency separation between peaks $(f_1 - f_2)$ increases with the coupling. $\dfrac{f_1 - f_2}{f} = K$ approximately for overcoupled high-Q circuits.

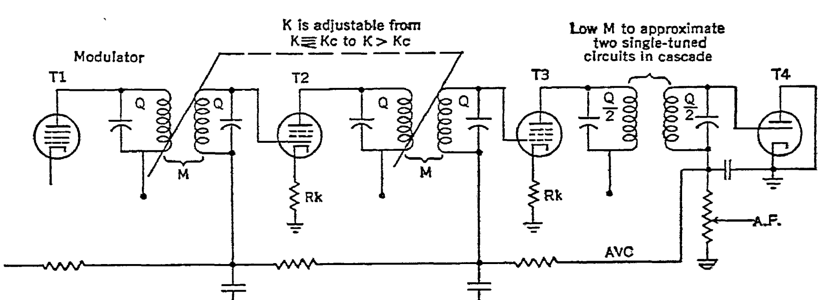


FIG. 3. I-f Amplifier Having Variable Selectivity

The valley between the peaks is filled in by the selectivity of a single-tuned circuit. A flat-topped selectivity curve can be approximated for all coefficients of coupling if the Q of each circuit of the coupled pair is equal to twice the Q of the single-tuned circuit. In practice, the i-f amplifier takes the form of Fig. 3, which consists of three double-tuned circuits. The coupling of the first two pairs is adjustable and capable of being overcoupled.

The coupling of the last pair is low so as to approximate the selectivity of two single-tuned circuits in cascade (one to fill in the valley of each overcoupled pair). The Q of the adjustable pairs is equal to twice the Q of the very loosely coupled pairs. The cathode resistors of about 100 ohms are unbypassed to minimize the detuning caused by variation of the bias by the avc. (See reference 3.)

Broad i-f amplifiers are sometimes used to reduce the need for frequency stability in the local oscillator in pushbutton sets or short-wave sets.

Tuning Stability. The amplifier should be tuned with enough capacitance (> 25 $\mu\mu$f) so that it will not be appreciably detuned by (*a*) replacement of tubes, (*b*) displacement of parts by vibration, (*c*) change in input capacitance of vacuum tubes caused by variation of gain. (This can be balanced by a small unbypassed cathode resistor (reference 3) of about 100 ohms; see Fig. 2.)

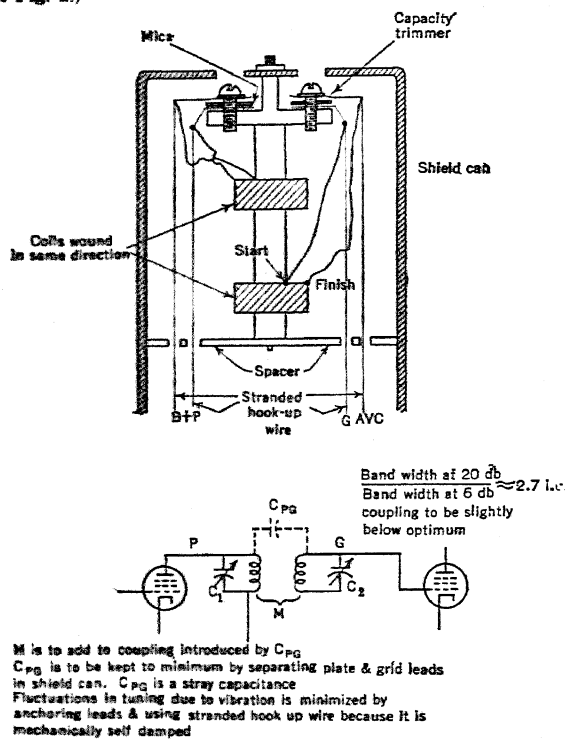


FIG. 4. Typical 0.455 Mc I-f Transformer

Where the amplifier will meet wide variations in temperature and humidity, the use of stable fixed condensers and tuning by means of adjustable iron-core coils is desirable.

The inductive coupling between windings can either aid or oppose the coupling due to the capacitance between the plate and grid terminals of the i-f transformer.

If very small transformers are desired, opposing inductive and capacitive coupling permits closer spacing of the windings. For coaxial coils, wound in the same direction, opposing coupling permits the coils to be partly self-shielded electrostatically by using the outside terminals as the low i-f potential terminals.

However, variation in the above capacitance, due to differences in production wiring or changing distance between wires due to vibration, causes much larger variations in the gain and bandwidth of an i-f stage with opposing couplings than with aiding coupling because the difference of two nearly equal quantities varies much more than their sum as either quantity is changed.

For the above reasons, whenever size and economy permit, aiding coupling is preferable to opposing coupling. However, the combined coupling should be sufficiently below optimum to prevent overcoupling due to production variations and vibration.

With either coupling, production will be much more uniform if the capacitive coupling is made as small as possible (see Fig. 4).

Feedback of I-f Harmonics. The second detector output must be well filtered and kept as far as possible from the r-f components so as to prevent harmonics of the intermediate frequency from being impressed on the r-f circuits where they can beat with the signal and cause "tweets."

I-F AMPLIFIERS FOR FM RECEIVERS. The same considerations hold for frequency modulation as for amplitude modulation except that:

1. Bandwidth must accommodate maximum signal frequency swing so as not to cause amplitude distortion.

2. Gain must be adequate to operate limiter on weakest signal.

3. Top should be reasonably flat so as not to overtax limiters and to reduce distortion in balanced detectors.

3*a*. Bandwidth at least 150 kc at −6 db. Adjacent channel attenuation ±400 kc at least 50 db.

4. Selectivity curve should be symmetrical.

5. Intermediate frequency for the 88–108 band is usually 10.7 Mc.

6. Amplifier consists usually of sets of double-tuned circuits.

7. A-m and f-m circuits usually are combined in one common shield can (but care must be taken to prevent interaction so that leads will not change couplings by adding capacity between windings).

8. To prevent detuning with avc, those tubes which have avc may be supplied with an unbypassed cathode resistor of about 100 ohms.

9. Tuning usually is accomplished by means of iron slug to increase the stability.

A typical design is given in Table 4 and Fig. 5.

Table 4. Typical Dual I-f Amplifier Stage for Am-Fm Set

	AM	FM
Frequency	0.455 Mc	10.7 Mc
$L_1 = L_2$ adjusted by iron core from	862 μh	4.2 μh
to	1395 μh	8.4 μh
Q	68–77	88
$\frac{K}{K_c}$	0.70 approx.	0.80
Gain	43 db	28 db
W_6	13.5 kc	215 kc
W_{20}	36 kc	540 kc
$\frac{W_{20}}{W_6}$	2.66	2.50

Note: The ratio of the bandwidths of two signals 20 db and 6 db stronger than the signal at resonance is a measure of the coupling. This is indicated as W_{20}/W_6 and is equal to 2.37 when $L_1C_1 = L_2C_2$, $Q_1 = Q_2$, and the circuits are critically coupled. When $Q_1 \neq Q_2$ as in a transformer feeding a diode, experience indicates that this ratio should be 2.7–3.0. When the transformer is used in the plate of a modulator tube, the ratio should be 2.5–2.6.

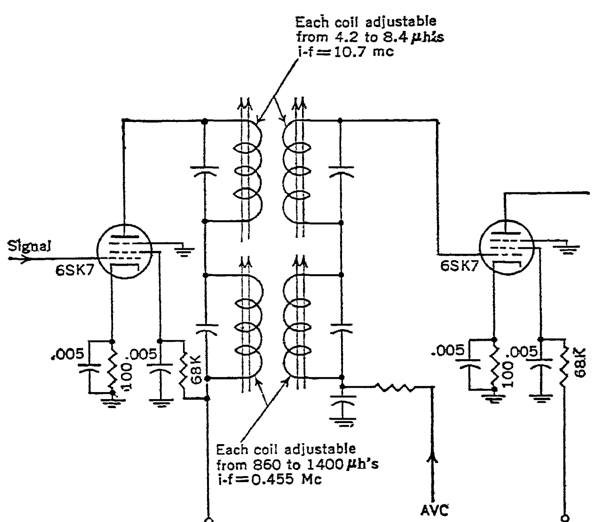


FIG. 5. Typical 0.455 and 10.7 Mc I-f Stage for Combined A-m and F-m Receiver

USEFUL RELATIONS FOR HIGH-Q CIRCUITS.

Selectivity can be expressed in two ways:

(a) With constant input as

$$S = \frac{\text{Output voltage at any frequency } f \text{ (or bandwidth } f_w \simeq 2[f - f_0])}{\text{Output voltage at the resonant frequency } f_0} \tag{1}$$

= a number less than unity

(b) With constant output as

$$A = \frac{1}{S} = \frac{\text{Input voltage at any frequency } f \text{ (or bandwidth } f_w \simeq 2[f - f_0])}{\text{Input voltage at the resonant frequency } f_0} \tag{2}$$

= a number larger than unity

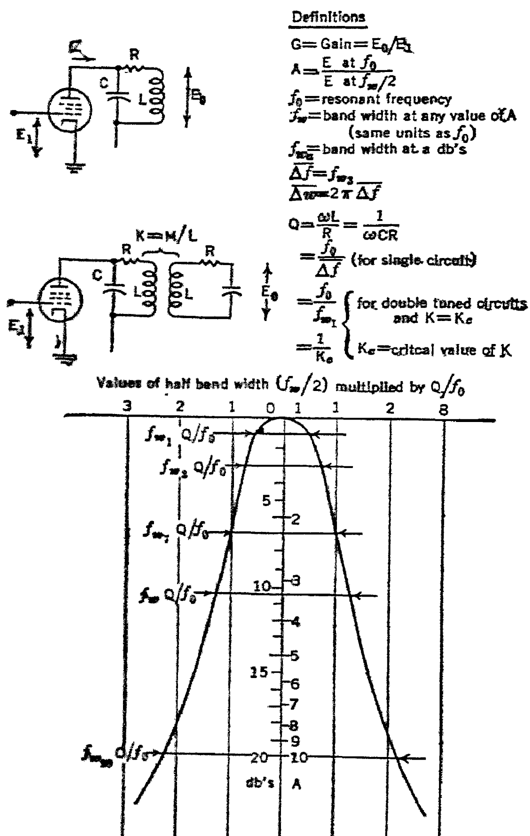


FIG. 6. Definition of Symbols in Text. Curve shows selectivity, A, of single double-tuned stage for $K = K_c$; and $Q_1 = Q_2$.

A. ***Single-tuned Stage*** (see Fig. 6 for definition).

1. $$Q = \frac{\text{Resonant frequency}}{\text{Bandwidth to signal 3 db greater than resonance}} = \frac{f_0}{\Delta f} \tag{3}$$

$\overline{\Delta f}$ is bandwidth for signal 3 db greater than the signal applied at the resonant frequency.

$$\overline{\Delta\omega} = 2\pi\,\overline{\Delta f} \tag{4}$$

2. *Selectivity* $= S = \dfrac{\text{Signal at } f \text{ (or } f - f_0)}{\text{Signal at } f_0} = \dfrac{1}{A}$ (5)

$$= \frac{1}{\sqrt{1 + 4Q^2\left[\dfrac{f - f_0}{f}\right]^2}}$$ (See Section 6, article 1, for universal selectivity curve) (5a)

3. *Gain per Stage*

$$G = g_m Q X \tag{6}$$

4. *Stage Gain Multiplied by Bandwidth at 3 db*
   (a) One stage

$$G\,\Delta\omega = g_m/C \tag{7}$$

   (b) n stages

$$G\,\overline{\Delta\omega} = \frac{0.835}{\sqrt{n}}\frac{g_m}{C} \text{ (approx.)} \tag{8}$$

B. *Double-tuned Stage* (Fig. 6).

The following formulas hold when $L_1C_1 = L_2C_2$; $Q_1 = Q_2$; K = coefficient of coupling; K_c = critical coupling coefficient.

1. When $K \leqq K_c$, then, for each stage,

(a) $$\frac{\text{Resonant frequency}}{\text{Bandwidth to signal 1 db greater than resonance}} = \frac{f_0}{f\omega_1} \geqq Q \tag{9}$$

(b) $$\frac{\text{Bandwidth to signal 20 db above resonance}}{\text{Bandwidth to signal 6 db above resonance}} = \frac{W_{20}}{W_6} \geqq 2.37 \tag{10}$$

2. *Selectivity*, S

(a) $K \leqq K_c$

$$S = \frac{\text{Signal at } f \text{ (or } f - f_0)}{\text{Signal at } f_0} = \frac{1}{A}$$

$$= \frac{1 + (K^2/K_c^2)}{\sqrt{\left[1 + (K^2/K_c^2) - 4Q^2\left(\dfrac{f - f_0}{f_0}\right)^2\right]^2 + 16Q^2\left(\dfrac{f - f_0}{f_0}\right)^2}} \tag{11}$$

(See Section 6, article 5.)

(b) $K = K_c$

$$S = \frac{1}{\sqrt{1 + 4Q^4\left(\dfrac{f - f_0}{f_0}\right)^4}} = \frac{1}{A} \tag{12}$$

3. *Gain per Stage*, G

(a) $K \leqq K_c$

$$G = g_m Q X \frac{K/K_c}{1 + (K^2/K_c^2)} \tag{13}$$

(b) $K = K_c$

$$G = \tfrac{1}{2} g_m Q X \tag{14}$$

4. *Stage Gain Multiplied by Overall Bandwidth*
   (a) One stage

$G\,\overline{\Delta\omega} = 0.707\,\dfrac{g_m}{C}$ where C is the actual tuning C and is $\tfrac{1}{2}$ of total C (approx.) (15)

   (b) n stages

$$G\,\overline{\Delta\omega} = \frac{1}{1.1\sqrt[4]{n}}\left(0.707\,\frac{g_m}{C}\right) = \frac{0.643}{\sqrt[4]{n}}\frac{g_m}{C} \tag{16}$$

13. WIDE-BAND I-F AMPLIFIERS

The amplification of video signals such as are used in television and pulse communications requires band widths wider than 200 kc. Hence techniques have been developed to design much wider band amplifiers of high overall gain. One special problem arising is that much of their use is for visually displayed information where the phase distortion introduced by a rectangular-topped selectivity curve cannot be tolerated. For this reason, the edges of the selectivity curve must be somewhat rounded.

ALTERNATIVE DESIGNS. Wide-band amplifiers may be designed using (1) synchronously tuned single-tuned circuits, (2) double-tuned circuits (loaded either on one or both sides, (3) stagger-tuned amplifiers, or (4) inverse-feedback amplifiers. Stagger-tuned amplifiers consist of n single-tuned circuits of poor skirt selectivity which are tuned to different frequencies so that the peak of one circuit tends to fill in a deficiency of the others. For instance a stagger-tuned pair consists of two single-tuned stages, one of which is peaked at a frequency higher than, the other lower than, the center frequency of the amplifier. In inverse-feedback amplifiers a fraction of the output of a single-tuned synchronous amplifier is fed back in degenerative phase to the input. More voltage, therefore, is fed back at resonance than off-resonance. Consequently, the overall gain is reduced more at resonance than off, and the nose of the selectivity curve is rounded.

FIGURE OF MERIT. The figure of merit of a wide-band amplifier is defined as

$$\text{Stage gain times overall bandwidth} = G\,\Delta\omega$$

where the band width $\Delta\omega$ means the band width for gains 3 db below the peak. The 3-db bandwidth is chosen because it makes the mathematics easier, it approximates the noise bandwidth of the receiver, and so facilitates signal-noise comparisons. Furthermore, with the usual coupling circuits, the rise time of pulses is quite simply connected with it. Defining rise time as the time required for the response to the step function to increase from 10 to 90 per cent of its final value, then

$$\text{Rise time} = \frac{0.7}{\Delta f}$$

Thus a 10-Mc (3 db down) i-f amplifier would have a minimum pulse rise time of 0.07 μs. (For a complete criterion of a pulse amplifier, percentage of pulse overshoot must also be considered.)

For a single-stage single-tuned circuit the gain of an amplifier is (Z is the plate load impedance)

$$G = g_m Z = g_m \frac{\omega^2 L^2}{r} = g_m Q\omega L = \frac{g_m}{\omega c}\frac{\omega}{\Delta\omega} = \frac{g_m}{C\Delta\omega}$$

so that

$$G\,\Delta\omega = g_m/C \tag{17}$$

which is also eq. (7). The factor C represents the total capacitance in the plate circuit, including tube output and input capacitances and the capacitance of circuit components to ground.

Since double-tuned circuits divide the total tube and circuit capacitance between two tuned circuits they use less capacitance per tuned circuit and have a higher figure of merit. For equal Q's this is $\sqrt{2}\,g_m/C$, and for one side loaded it is $2g_m/C$.

SYNCHRONOUS SINGLE-TUNED CIRCUITS.* The simplicity and stability of an amplifier made up of single-tuned circuits all tuned to resonance commend it to the designer, and most 2- to 3-Mc-wide amplifiers are of this type. For really wide-band high-gain amplifiers the decrease in bandwidth arising from multiplication of the successive selectivity curves makes this type unsuitable. Table 5 shows the overall bandwidth of n-stage amplifiers in terms of the stage bandwidth. It shows that a nine-stage synchronous single-tuned amplifier with an overall bandwidth of 4 Mc requires a single-stage bandwidth of 14.3 Mc with its correspondingly low gain. It can be shown that the maximum overall bandwidth of synchronous single-tuned circuits occurs for a mean stage gain of

$$G = \sqrt{\epsilon} = 4.34 \text{ db}$$

Table 5. Overall Bandwidth of Sychronously Tuned Amplifiers

Stages	% of Stage Bandwidth
2	64
3	51
4	44
5	39
6	35
7	32
8	30
9	28

DOUBLE-TUNED CIRCUITS. Double-tuned circuits have a theoretical advantage in the figure of merit. However, the large number of adjustments (three per stage) and the criticalness of some of them (interlocking effects) make design and maintenance most difficult. For this reason double-tuned circuits have been used very little.

STAGGER-TUNED SINGLE-TUNED CIRCUITS. Wallman defines an exact staggered pair (or triple) of bandwidth Δf and geometrically centered at f_0 as consisting of (a) two single-tuned circuits of dissipation factor $d = 1/Q$ and frequencies αf_0 and f_0/α, and

* The following three paragraphs were abstracted from a paper by H. Wallman, "Stagger-tuned i-f Amplifiers," *Rad. Lab. Report* 524 NDRC Div. 14.

(for one triple) (*b*) one single-tuned circuit of dissipation factor $\delta = \Delta f/f$ centered at f_0. d and α are plotted as functions of $\delta = \Delta f/f$ in Figs. 7 and 8. (*Note*: The values of d and α for a pair are not the same as their value for a triple.)

The stage gain times overall bandwidth of (*a*) a single-tuned stage, (*b*) an exact staggered pair, and (*c*) an exact staggered triple is

$$G\,\overline{\Delta\omega} = g_m/C$$

However, the same 3-db bandwidth $\Delta\omega$ is obtained in the first case (*a*) for one tuned circuit; in the second case (*b*) for two tuned circuits, and (*c*) for three tuned circuits.

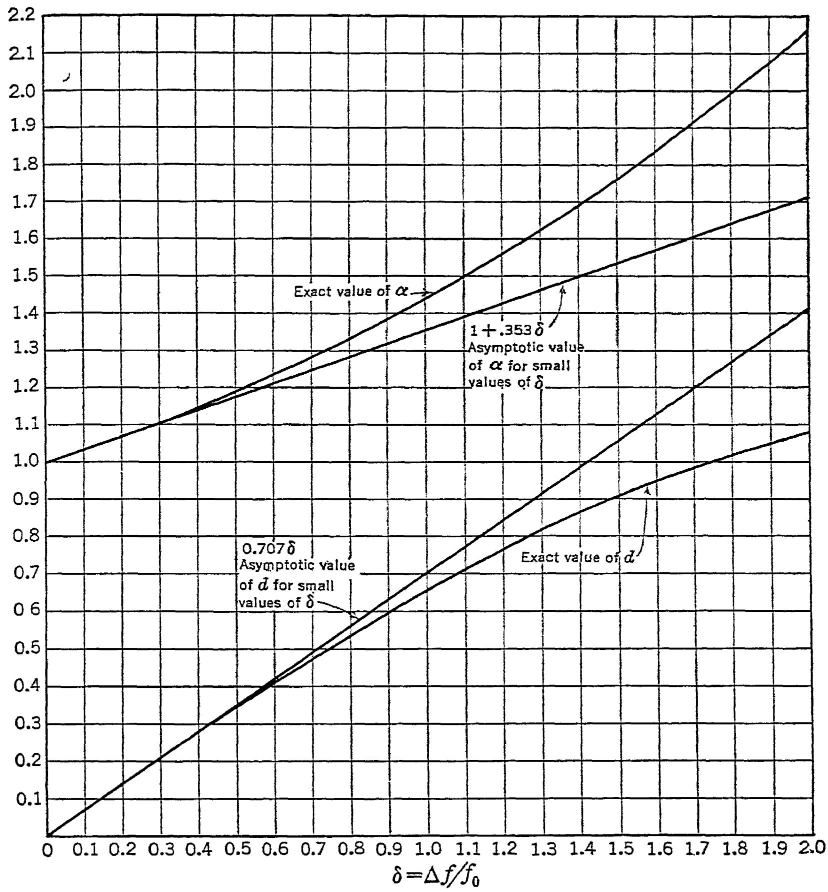


FIG. 7. Design Curves for an Exact Flat Staggered Pair

The algebra in the analysis of these circuits is quite involved, but Wallman has reduced the design to charts. Given the tube type and the general layout, the g_m is known and an estimate of C is made. This determines the product of *single*-stage gain and *overall* bandwidth, i.e., $G\,\overline{\Delta\omega} = g_m/C$ of the pair or triple.

The overall bandwidth of the amplifier can be obtained from the following approximate relations

1. $$\text{Bandwidth of } n \text{ pairs} = \frac{\text{Bandwidth of one pair (i.e., } \overline{\Delta\omega})}{1.1\sqrt[4]{n}}$$

2. $$\text{Bandwidth of } n \text{ triples} = \frac{\text{Bandwidth of one triple (i.e., } \overline{\Delta\omega})}{1.06\sqrt[6]{n}}$$

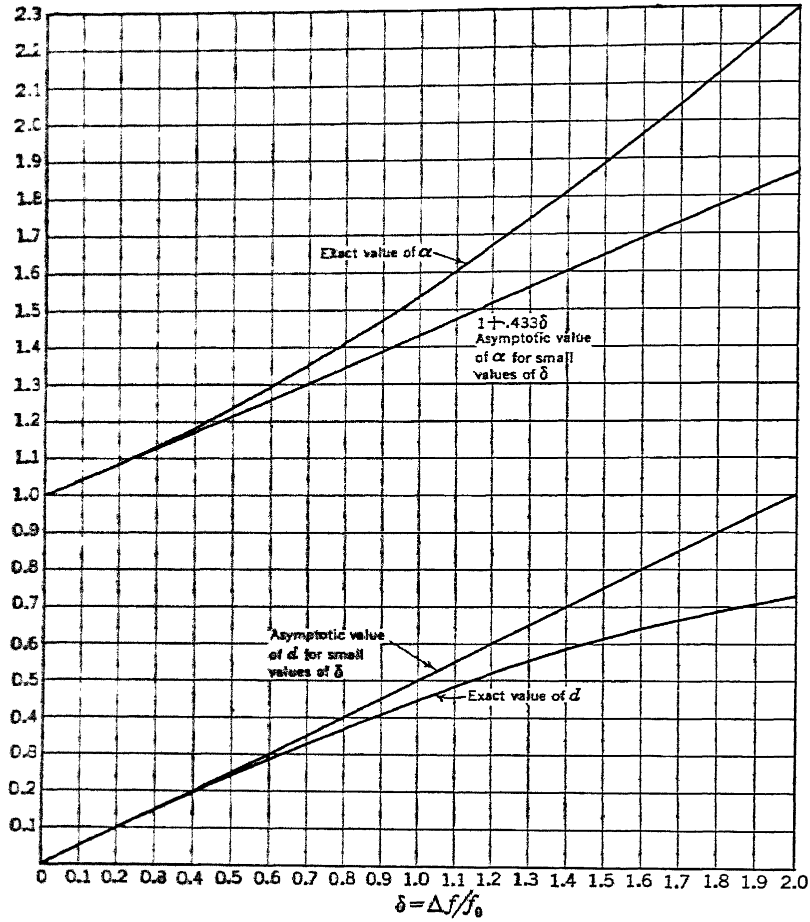


FIG. 8. Design Curves for an Exact Staggered Triple

Conversely, the overall bandwidth and overall gain being known, the number of stages as well as the required unit performance can be determined.

Example. A nine-stage amplifier having a gain of 108 db centered on 30 Mc is wanted. 6AC7's having a g_m of 0.0085 ampere/volt and a total capacitance of 25.5 $\mu\mu f$ are to be used.

Nine stages divide conveniently into three triples.

The stage gain is 108/9 = 12 db, that is, a voltage gain of 4 times.

$$G\,\overline{\Delta\omega} = \frac{g_m}{C} = \frac{0.0085}{25.5} \times 10^{12} = 332 \times 10^6$$

$$\overline{\Delta f} = \frac{332 \times 10^6}{4 \times 2\pi} = 13.3 \text{ Mc}$$

Each triple has a bandwidth of 13.3 Mc. The three triples that make up the amplifier have an overall bandwidth of

$$\text{Overall bandwidth} = \frac{13.3}{1.06\sqrt[6]{3}} = 10.55 \text{ Mc}$$

Constants making up each triple are found from Fig. 8 as follows:

$$\delta = \frac{\Delta f}{f_0} = \frac{13.3}{30} = 0.443$$

$$d = \frac{1}{Q} = 0.215 \quad \text{and} \quad \alpha = 1.21; \quad Q = 4.65$$

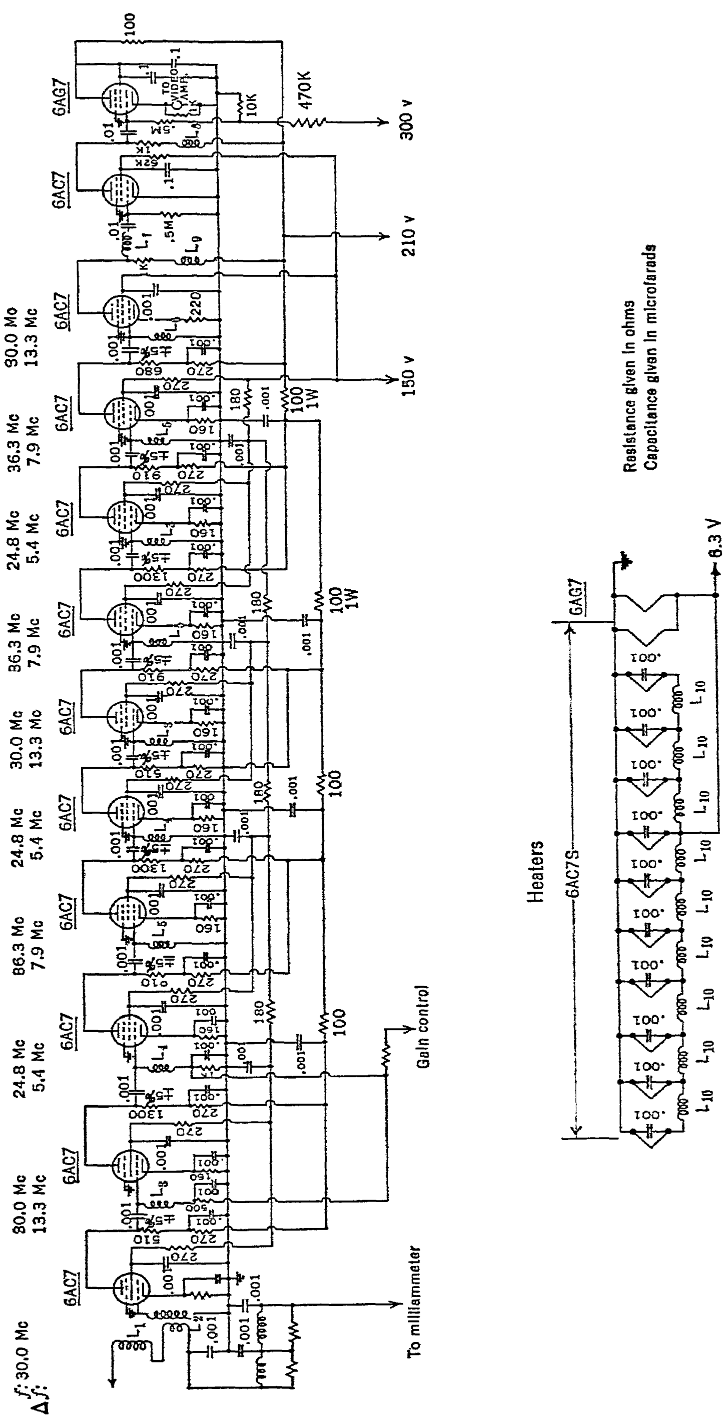


FIG. 9. Wide-band Amplifier Consisting of Three Staggered Triples, Having an Overall Gain of 108 db and an Overall Bandwidth of 10.5 Mc

The amplifier, therefore, is made up of three triples, each of which consists of:

1. One circuit tuned to $30 \times 1.21 = 36.3$ Mc and having a Q of 4.65.
2. One circuit tuned to $30/1.21 = 24.8$ Mc and having a Q of 4.65.
3. One circuit tuned to 30 Mc and having a Q of $1/0.443 = 2.26$.

Since $Q = \omega CR$ (R in this case is the equivalent resistance across the tuned circuit), then

$$R = \frac{Q}{\omega C} = \frac{4.65}{2\pi \times 36.3 \times 10^6 \times 25.5 \times 10^{-12}} = 790\omega \text{ for the 36.3-Mc circuit}$$

$$R = \frac{4.65}{2\pi \times 24.8 \times 10^6 \times 25.5 \times 10^{-12}} = 1150\omega \text{ for the 24.8-Mc circuit}$$

$$R = \frac{2.26}{2\pi \times 30 \times 10^6 \times 25.5 \times 10^{-12}} = 470\omega \text{ for the 30-Mc circuit}$$

The amplifier is shown in Fig. 9. Actually the physical values of the shunting resistors are higher than calculated because of the effect of tube loading and of the finite Q of the coils which effectively shunt the physical resistor to the calculated value. The final value of these resistors must be determined experimentally.

The bandwidth of each stage of the triple is determined from the relation

$$G\,\overline{\Delta\omega} = \frac{g_m}{C}$$

$$g_m R\,\overline{\Delta\omega} = \frac{g_m}{C}$$

$$\overline{\Delta\omega} = \frac{1}{RC}$$

$$\overline{\Delta f} = \frac{1}{2\pi RC}$$

so that the results for each triple can be tabulated as follows:

f_0, Mc	Q	$\overline{\Delta f}$, Mc	Shunt Resistance, ohms
(1) 36.3	4.65	7.9	790
(2) 24.8	4.65	5.4	1150
(3) 30.0	2.26	13.3	470

Wallman gives tables which extend the design to quadruples, and so forth, called n-upples.

TYPICAL INVERSE FEEDBACK I-F AMPLIFIER CIRCUITS. Typical feedback circuits are shown in Figs. 10a, 10b, and 10c. These are drawn without regard to d-c potentials. Thus the grids must be protected from the plate voltages by blocking condensers or, preferably, transformers having unity coupling.

The feedback chain of Fig. 10a is of little use, despite its high gain times bandwidth factor, because its gain cannot be controlled. Figures 10b and 10c show respectively a

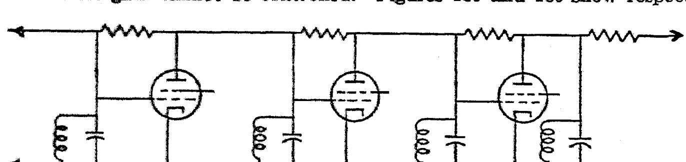


FIG. 10a. Ideal Inverse Feedback Chain

typical feedback pair and feedback triple. In each case the voltage which controls the gain is applied to the first tube.

A complete i-f amplifier using 6AK5's and inverse feedback is shown in Fig. 11.

This amplifier consists of two triples made up respectively of tubes V-301, V-302, V-303, and V-304, V-305, V-306. Gain control is applied to V-301 and V-304. This amplifier is centered on 60 Mc and has an overall gain of approximately 90 db and a bandwidth at 3 db of approximately 10 Mc.

The design of feedback amplifiers is not as readily computed as that of stagger-tuned amplifiers and therefore is based to a greater extent on test and experience.

For an excellent study of feedback amplifiers see reference 9, 10, 11, 13, and 14.

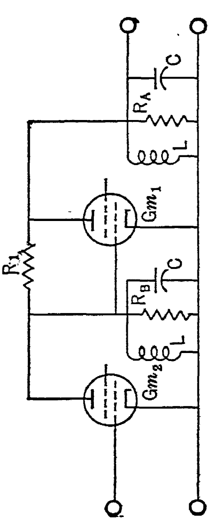


Fig. 10b. Inverse Feedback Pair. The gain is controlled by biasing the first tube because it is not involved in the feedback loop.

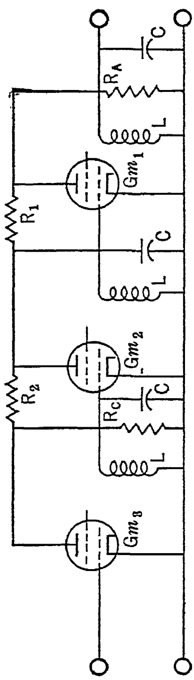


Fig. 10c. Inverse Feedback Triple. The gain is controlled by biasing the first tube because it is not involved in the feedback loop.

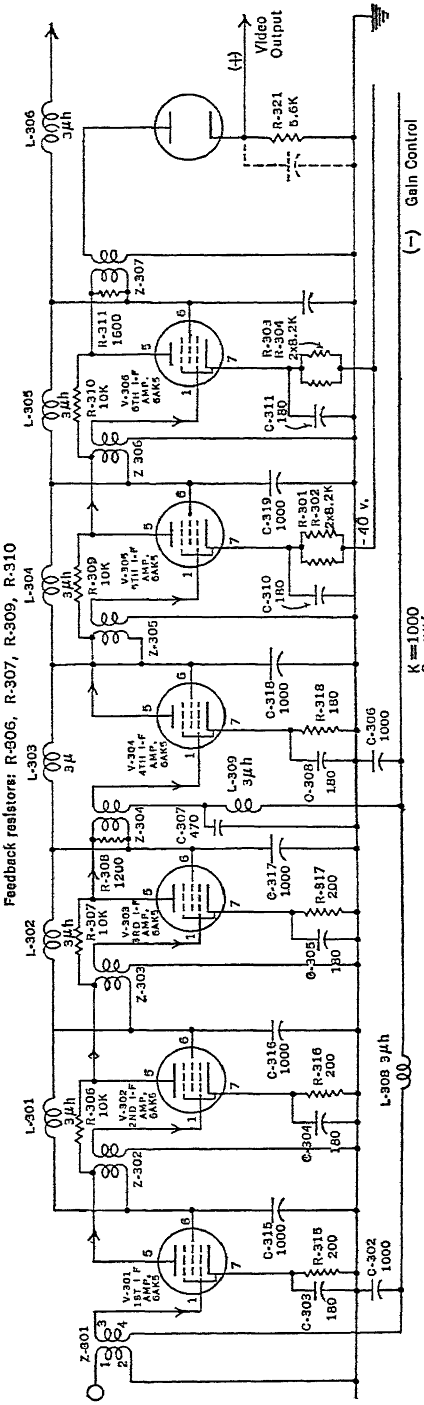


Fig. 11. 60-Mc Inverse Feedback I-f Amplifier Consisting of Two Triples

STAGGER-TUNED AMPLIFIERS VS. INVERSE FEEDBACK AMPLIFIERS. The choice between a stagger-tuned and an inverse-feedback amplifier depends largely on the circumstances surrounding its use. There is no fixed answer as to which is best for all conditions. The differences are outlined below.

Figure of Merit. Both these types of amplifiers have been shown to have a much higher figure of merit than the synchronous single-tuned circuit type. Theoretically the inverse-feedback amplifier can have a higher one than the stagger-tuned, but it is difficult to achieve this in practice, so that they may be assumed equal.

Gain Control. Gain adjustment may be desired (1) to change the output, (2) to change the gain with time (gain time control, gtc in radar applications), (3) to reduce interference, (4) for automatic gain stabilization (ags), (5) for automatic gain control (avc).

A stagger-tuned amplifier's bandwidth being practically independent of the tube's g_m, gain control can be applied to any tube.

An inverse-feedback amplifier can use for gain control only those tubes which are not in the feedback chain. Otherwise, the bandwidth will change with gain.

Gain Variation with g_m. The gain varies more slowly with g_m in a feedback amplifier (approximately as $g_m^{1/2}$) than in others (linearly with g_m). This means that the gain of amplifiers not equipped with automatic gain stabilization (ags) will be less susceptible to voltage change, tube ageing, and tube replacement if of the feedback type.

Effect of Replacement Tubes on Bandwidth. In a feedback amplifier, the resonant frequency, and therefore the overall bandwidth, is less affected by replacement tubes of differing capacitance. However, the bandwidth is more affected by tubes of differing g_m.

Resonant Frequency. An inverse-feedback amplifier has all its circuits tuned to the center frequency of the pass band.

Independence of Tuning. In a stagger-tuned amplifier the tuning of one stage is independent of the tuning of the other stage.

BIBLIOGRAPHY

1. Dean, C. E., Band Width Factors for Cascade Tuned Circuits, *Electronics*, July 1941.
2. Spaulding, F. E., Jr., Design of Superheterodyne Intermediate-frequency Circuits, *RCA Rev.*, April 1940.
3. Freeman, R. L., Use of Feedback to Compensate for Vacuum Tube Input Capacitance Variation with Bias, *Proc. I.R.E.*, November 1938.
4. Wheeler, H. A., and J. K. Johnson, High Fidelity Receivers with Expanding Selectors, *Proc. I.R.E.*, June 1935.
5. Wheeler, H. A., Wide Band Amplifiers, paper presented at joint meeting of I.R.E. and A.I.E.E. in New York, April 18, 1945.
6. Foster, D. E., and J. A. Rankin, Intermediate-frequency Values for Frequency-modulated Wave Receivers, *Proc. I.R.E.*, October 1941.
7. Zworykin, V. K., and G. A. Morton, *Television*, Chapter 17. John Wiley (1940).
8. Wallman, Henry, Stagger-tuned I-f Amplifiers, Radiation Laboratory Report 524, NDRC Div. 14.
9. The Theory of I-f Amplifiers with Negative Feedback, Canadian National Research Council Radio Branch Report PRA-59 (1942).
10. Beveridge, H. N., Information on Broad-band Feedback I-f Amplifiers, CRG Report 93, Naval Research Laboratory, Oct. 22, 1945.
11. Abraham, W., Equations for Feedback Amplifiers, CRG Technical Memorandum 217, Nov. 26, 1945.
12. Wide Band Amplifiers for Television, *Proc. I.R.E.*, July 1939.
13. Valley and Wallman, *Vacuum Tube Amplifiers*, Rad. Laboratory Series No. 18, Chapters 4, 5, and 6, McGraw-Hill.
14. Lebenbaum, M. T., Intermediate Frequency Amplifiers, *Very High Frequency Techniques*, Radio Research Laboratory Staff, Chapter 33, Vol. II, McGraw-Hill.

MODULATORS

By J. E. Young

Modulation is the process whereby the amplitude, or other characteristic, of a wave is varied as a function of the instantaneous value of another wave. The first wave, which is usually a single-frequency wave, is called the "carrier wave"; the second is called the "modulating wave."

The carrier frequency is the frequency of the carrier wave.

Sidebands are the frequency bands on each side of the carrier frequency within which fall the frequencies of the waves produced by the process of modulation. In amplitude modulation, the width of a transmitted sideband is usually no greater than the bandwidth of the modulating wave. The bandwidth of a frequency-modulated signal is determined by both the modulating frequency and the frequency swing employed.

Percentage modulation (amplitude modulation) is the ratio of the difference between the maximum current (I_m) when a signal is impressed and the maximum unmodulated current (I_s) to this maximum unmodulated current, expressed in percentage, or $M = 100(I_m - I_s)/I_s$. For less than 100 per cent modulation this may optionally be expressed as the ratio of half the difference between the maximum and minimum amplitudes of a modulated wave to the average amplitude, expressed in percentage.

Percentage modulation (frequency modulation) is the ratio of the frequency swing when a signal is impressed, to an arbitrary frequency swing which is defined as 100 per cent modulation, expressed in percentage. The frequency swing defined as 100 per cent modulation is different for various services. For f-m broadcasting, for example, 75 kc is used.

14. TYPES OF MODULATION

A single-frequency current wave can be expressed as

$$i = A \sin (\omega t + \theta) \tag{1}$$

where A is the amplitude, $\omega/2\pi$ the frequency (when constant), and θ the relative phase. If any of the three independent magnitudes, A, ω, or θ, is slowly varied (slow in comparison to $\omega/2\pi$) the wave is said to be modulated. The three cases are called amplitude, frequency, and phase modulation, respectively.

If ω and θ are held constant but A is varied sinusoidally, so that $A = A_0(1 + m \sin \omega_1 t)$, eq. (1) becomes

$$i = A_0(1 + m \sin \omega_1 t) \sin (\omega t + \theta) \tag{2}$$

which, when expanded trigonometrically, gives

$$i = A_0 \sin (\omega t + \theta) + \frac{mA_0}{2} \{\sin [(\omega + \omega_1)t + \theta] + \sin [(\omega - \omega_1)t + \theta]\} \tag{3}$$

The quantities in the bracket of the second term on the right represent sum and difference frequencies of the two original frequencies. They are the sidebands. Comparison of these two equations shows that a wave of a single frequency and periodically varying amplitude is mathematically equivalent to a wave of constant amplitude and frequency and a pair of sidebands.

Roder has shown (*Proc. I.R.E.*, Vol. 19, 2145 [1931]) that, if A and θ are constant but $\omega = \omega_0(1 + k_f \cos \mu t)$, the current is (let $\omega_0 t = \omega t + \theta$)

$$\begin{aligned} i = A_0 \sin \omega_0 t &+ J_1(m_f)[\sin (\omega_0 + \mu)t - \sin (\omega_0 - \mu)t] \\ &- J_2(m_f)[\sin (\omega_0 + 2\mu)t - \sin (\omega_0 - 2\mu)t] \\ &+ J_3(m_f)[\sin (\omega_0 + 3\mu)t - \sin (\omega_0 - 3\mu)t \end{aligned} \tag{4}$$

where $m_f = k_f\omega/\mu$ so that m_f is the ratio between the maximum frequency shift and the audio frequency; also $J_n(m)$ means the Bessel function (see Section 1, article 14) of the first kind and nth order for the argument m. In this expression there are theoretically an infinite number of sidebands, although the amplitude of all those of higher order than the first is usually negligible.

If A and ω are constant but $\theta = \theta_0(1 + K_p \sin \mu t)$, Roder gives the current as identical in form with eq. (4) except that $m = K_p\theta_0$. Here, again, there are theoretically an infinite number of sidebands, but only those of the first order are of importance or of appreciable amplitude.

Except for some very early and ineffective attempts to use frequency or phase modulation, amplitude modulation was used exclusively for communication and broadcasting up to 1936. Exploration into the higher radio frequencies and the adaptation of these frequencies to wide use for numerous services has brought frequency- and phase-modulation techniques into great importance, since the wider frequency bands required, if they are to be used effectively, become available as the usable frequency spectrum is extended into the hundreds and thousands of megacycles.

AMPLITUDE MODULATION. Since the e^2 term of the power series of development for current in a non-linear circuit gives rise to the term $\sin \omega_1 t \sin \omega t$, characteristic of modulation, it follows that modulation occurs *whenever any of the circuit parameters vary with instantaneous voltage* (or current). In particular, if it is assumed that a single high-frequency voltage $e_p' = E_s' \cos (\omega_s t + \theta_s)$ is introduced in the plate circuit of a vacuum tube and a varying audio voltage $e_g = \sum_k E_{gk} \cos (\omega_k t + \theta_k)$ is impressed on the grid, then, from Section 5, article 21, eq. (17), the terms having frequencies near that of $\omega_s/2\pi$ are

$$i_{ps} = \frac{E_s \cos(\omega_s t + \theta_s')}{z_s'} - \frac{r_p}{2}\frac{\partial r_p}{\partial e_p}\sum_k \frac{\mu E_s E_k}{z_s' z_k'}\left[\frac{\cos(\omega_s t + \omega_k + \beta_{4k})}{z'_{(s+k)}} + \frac{\cos(\omega_s t - \omega_k t + \beta_{5k})}{z'_{(s-k)}}\right] \quad (5)$$

where the phase angles are developed as illustrated there in eq. (22). This may optionally be written (assume the phase angles of $z_{(s+k)}$ and $z_{(s-k)}$ negligible as is usual in practical circuits).

$$i_p = \frac{E_s \cos(\omega_s t + \theta_s')}{z_s'}\left[1 - r_p \frac{\partial r_p}{\partial e_p}\sum \frac{\mu E_k \cos(\omega_k t + \theta_k')}{z_k' z_s'}\right] \quad (6)$$

Then the percentage modulation is

$$M = -\frac{100 r_p}{z_s'}\frac{\partial r_p}{\partial e_p}\left[\sum_k \frac{\mu E_k \cos(\omega_k t + \theta_k)}{z_k'}\right]_{\max} \quad (7)$$

The expression in the bracket is simply the maximum value of the audiofrequency plate current. These currents are shown in Fig. 1.

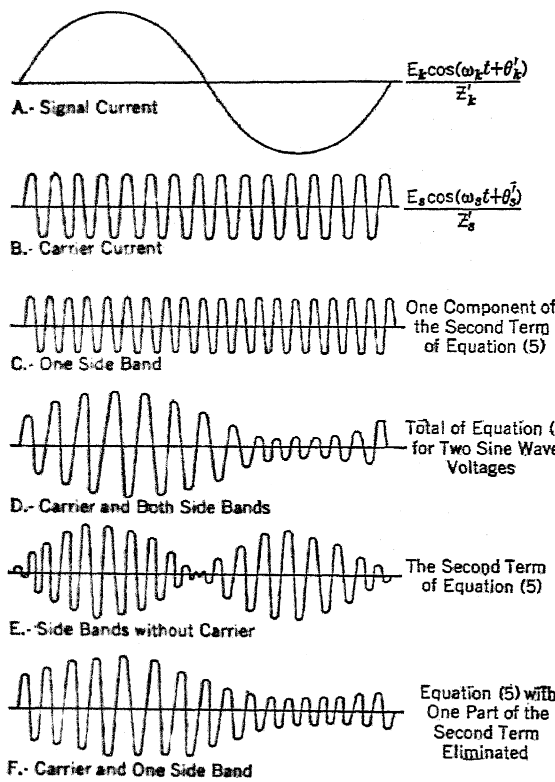


FIG. 1. Combination of Sine Wave Carrier and Signal Currents

METHODS OF PRODUCING AMPLITUDE MODULATION. Systems of modulation may be divided into two classes:

1. Systems in which the impedance of a r-f oscillator, amplifier, or combination of amplifiers is varied by the modulating wave.

2. Systems employing a constant-impedance r-f oscillator, or amplifier, having a variable impedance in series with it in which the variation of the magnitude of the series

impedance is controlled by the modulating wave, thereby controlling the input to the r-f unit.

The first method of modulation may be accomplished by varying the grid bias of the r-f oscillator or amplifier if a triode is used or, alternatively, varying the suppressor or screen-grid potential if multielectrode tubes are used.

The second method depends for its operation on the introduction of the modulating emf in series with the plate power supply of the modulated oscillator or amplifier. Since appreciable amounts of power are required to accomplish modulation in this manner, auxiliary tubes are employed as modulators.

15. GRID MODULATION

The term *grid modulation* is commonly used to describe modulation systems of the first class mentioned above. Modulation is effected by varying the d-c grid bias voltage or suppressor grid voltage at an a-f rate, by connecting the modulating source, such as the output transformer of an audio-amplifier stage, between the bias source and the r-f stage. Since the grid input resistance is high, the audio amplifier need be capable of only a few watts output. The power required is still further reduced in some designs by the use of a tube in the r-f circuit which is capable of large plate currents without exceeding zero grid bias. In this case the a-f amplifier looks into a circuit of substantially infinite impedance. The operating bias for the r-f stage is determined as follows: The r-f tube is first biased to cutoff, with no r-f grid voltage. The r-f grid voltage is then increased until the knee of the saturation curve is reached. Then, with a constant r-f grid voltage, the d-c grid voltage is increased until the plate current is substantially zero. If e_1 is the d-c grid voltage for maximum output, and e_2 is the d-c grid voltage for zero output, the operating grid voltage should be $(e_2 + e_1)/2$. The peak a-c voltage required from the audio stage to produce complete modulation is $(e_2 - e_1)/2$. Because of the relatively low efficiency obtained, grid modulation methods are generally applied only where the amount of modulated power required is very small. Higher powers are sometimes obtained by following the grid modulated stage by r-f amplifiers. The outstanding advantages of grid modulation are its simplicity and cheapness, which make it attractive for small, light-weight applications, such as portable or airborne.

POWER AND EFFICIENCY OF GRID-BIAS-MODULATED AMPLIFIERS. The following discussion may be applied in determining peak power requirements, efficiency, driving power, etc., of class B r-f amplifiers as well as bias-modulated amplifiers, since the tubes operate under the same conditions in each case. The peak power output capability of the tube must be four times the carrier power if complete modulation is to be obtained. This follows from the fact that the output current and voltage must be doubled on peaks of 100 per cent modulation. The efficiency is proportional to the percentage of modulation, and it is of the order of 20 to 33 per cent for normal carrier or unmodulated conditions. The ratio of driving power of the r-f exciter to carrier power of the modulated stage should not exceed 1 : 10, since the exciter must maintain a constant r-f voltage on the grid of the modulated tube. The following formulas show the current, voltage, power relations, and effect of the variation of the percentage of modulation.

Let W_P = peak instantaneous power output; W_C = carrier output power; W_A = average power output, 100 per cent modulation (used in determining heating effect); i_{pc} = peak plate current averaged over an r-f cycle, 100 per cent modulation; i_{pm} = average plate current, unmodulated; f_p = maximum instantaneous plate efficiency averaged over an r-f cycle, 100 per cent modulation; f_c = average plate efficiency, carrier unmodulated; F = modulation factor = $M/100$; W_{LP} = plate loss averaged over an audio cycle, 100 per cent modulation; and W_{LC} = average plate loss, carrier unmodulated.

Assume that d-c plate voltage remains constant, and that modulation is effected by variation of plate current and efficiency. Then:

$$F = \sqrt{\frac{W_P}{W_C}} - 1 \tag{8}$$

$$W_A = 1.5W_C \tag{9}$$

$$i_{pm} = 2i_{pc} \tag{10}$$

$$f_p = f_c(1 + F) \tag{11}$$

$$W_{LP} = W_{LC} \tag{12}$$

16. PLATE MODULATION

The earliest system of plate modulation, devised by Heising, was known as the constant-current system. See Fig. 2. It derived its name from the fact that variation of the power input to the r-f stage is obtained by using another tube or tubes as an absorber. The current drawn from the supply source remains constant and shifts between the absorber and useful load as the absorber or modulator impedance is changed by the a-f voltage impressed in its grid. Because of the curvature in the extremes of the vacuum-tube characteristics when used as a modulator, the percentage of modulation, where the r-f stage and modulator are both supplied with the same plate potentials, is limited. To obtain complete modulation, the d-c voltage impressed on the radio stage is often reduced by a series resistor. The a-f voltage developed by the modulator is transferred, unattenuated, by a suitable by-pass capacitor connected around the resistor. It is usually possible to obtain complete modulation by reducing the r-f amplifier voltage to 60 or 70 per cent of the modulator voltage.

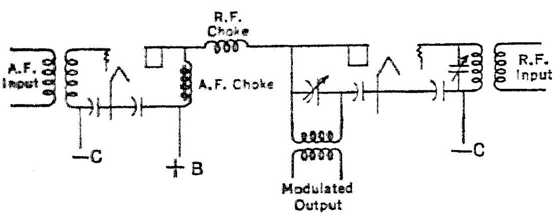


Fig. 2. Heising Amplitude Modulation

The high efficiency of plate-modulated class C amplifiers has led to their extensive use in high-power telephone transmitters. Modulator efficiency has been greatly improved by matching the modulator to the class C amplifier load by means of a coupling transformer and by using pairs of modulator tubes in class B circuits. Radio-frequency amplifiers following the modulated stage may be dispensed with entirely in this system, and power outputs obtained which are limited only by the capabilities of the vacuum tubes used in the modulator and modulated amplifier, with high overall efficiency. For the calculation of such circuits the reader is referred to p. 7-16. Transformer-coupled modulator circuits may be calculated by treating the modulator as a transformer-coupled a-f amplifier terminated on the secondary side of the coupling transformer by a load resistance equal to the modulated amplifier d-c plate voltage divided by its plate current. It should be noted that, where transformer-coupled modulators are used, the current delivered by the power source is no longer constant. In fact, in the case of the pushpull modulator, the current required by the modulator is very largely second harmonic. Special attention must be paid to the filter design, since harmonic voltage developed across the power supply terminals

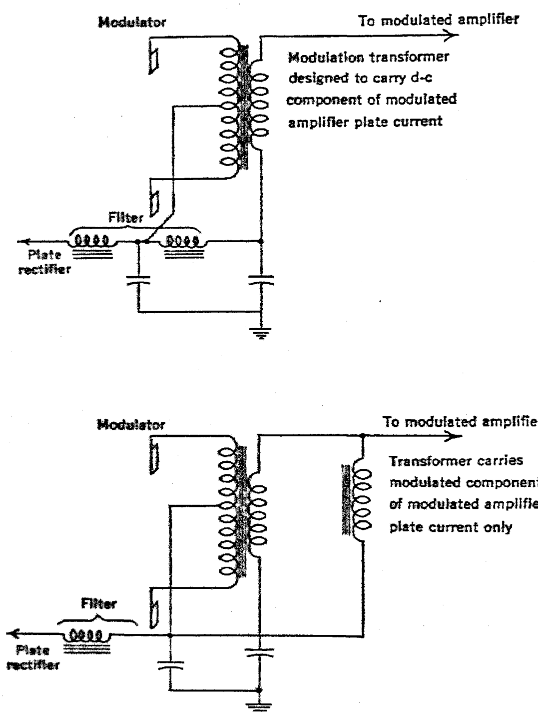


Fig. 3. Modulator-amplifier Coupling Circuits

by the modulator current demand will be applied to the modulated amplifier and appear as distortion in the modulated envelope. Recommended coupling circuits are shown in Fig. 3.

The modulated amplifier is operated class C; however, if distortion is to be minimized, special considerations are involved. It will be noted that the applied plate supply voltage to the class C stage has the wave shape of the modulating wave. It follows that, for 100 per cent modulation, the voltage rises to twice the carrier level, and the plate current must likewise rise to twice its quiescent level. Adequate grid excitation must be provided to insure this increase in plate current. This is usually achieved by driving the modulated amplifier somewhat harder than is required to produce rated carrier power at maximum efficiency. In addition, grid leak bias is used, and since the grid current decreases as the plate voltage increases (and the plate robs the grid of some of its electrons) the grid bias falls, the amount of grid driving power falls, and, owing to the combination of decreased bias and driver regulation, the a-c grid driving voltage increases, thus producing a condition favorable to a linear relation between modulating plate voltage and r-f output voltage.

The rise in driving voltage required to minimize distortion is, to some extent, a function of the design of the tube. Some types of tubes require less increase in drive, as the plate voltage rises, than is obtained by the combination of grid leak bias and driver regulation. Lowest distortion will be obtained in these cases if a combination of fixed and grid leak bias is used. The desired combination may easily be obtained by by-passing part of the grid leak with a capacitor large enough to retain most of its charge over the lowest a-f cycle.

Another method of obtaining the necessary variation in drive is to modulate the driver stage. It is usually not necessary, for optimum results, to use a depth of modulation greater than 30 or 40 per cent for this stage, when the modulated amplifier is modulated 100 per cent.

Tetrodes and pentodes may be used as modulated amplifiers. If low distortion is required, it is necessary to modulate the screen-grid voltage.

As in driver modulation, it is usually not necessary to modulate the screen voltage completely. The modulating voltage may be obtained from the plate circuit of the tube through a blocking capacitor and dropping resistor or, more conveniently, by a tertiary winding on the modulation transformer.

Let F = modulation factor; E_M = maximum plate voltage averaged over an r-f cycle; E_C = d-c plate voltage; W_{LP} = watts output, complete modulation; W_{LC} = watts output, normal carrier; W_M = modulator output power; and W_{CI} = modulated amplifier input power.

The following formulas show the current voltage relations in a plate-modulated amplifier:

$$E_M = E_C(1 + F) \tag{13}$$

$$W_{LP} = 1.5W_{LC} \tag{14}$$

$$F = \sqrt{\frac{2W_M}{W_{CI}}} \tag{15}$$

The last equation assumes that the modulated amplifier matches the output impedance of the modulator either directly or through a transformer, since, owing to limitation of the plate swing of the modulator, incorrect matching may limit the modulation factor, even though the modulator is potentially capable of producing complete modulation.

17. COMPARISON OF MODULATION SYSTEMS

Any comparison of modulation systems develops into a comparison of means of getting a specified amount of modulated power into a load, or antenna, under specified conditions of operation, weight, portability, fidelity, etc. For light-weight, portable transmitters, grid modulation has some advantage, although it is not capable without feedback of high fidelity. For most other applications plate modulation is used either directly, in the last r-f amplifier, or in a stage followed by one or more r-f amplifiers. In the last case, such amplifiers are usually of the high-efficiency Doherty type. For transmitters operating in the standard broadcast band where a transmission is almost always over a single path, both the "high-level" system of modulation in the last r-f amplifier, and the "low-level" system using Doherty amplifiers, are satisfactory. However, where multipath transmission exists, and incidental phase modulation may cause objectionable distortion, high-level modulation is generally preferred.

FREQUENCY MODULATION. Equation (4), in Article 14 was shown to represent the modulated r-f current for either the frequency-modulated wave or the phase-modulated wave. It will be noted, however, that the frequency swing, in frequency modulation, depends on the amplitude of the modulating signal but is independent of its frequency, whereas, in phase modulation, the resulting frequency swing is proportional to both the amplitude and the frequency of the modulating signal. It is possible, therefore, to convert from frequency modulation to phase modulation or vice versa by simply operating on the frequency-amplitude characteristic of the modulating signal. Frequency modulation is more generally used than phase modulation in broadcasting and communication systems. However, because of the ready conversion from one system to the other, the basic modulator employed in the transmitter is sometimes a phase-modulated device, which by alteration of the modulating-signal frequency characteristic is converted to produce frequency modulation. The relation between phase shift, frequency swing, and modulating signal frequency is given by:

$$\theta = \frac{K_f \omega}{\mu}$$

where θ = phase shift in radians.
$K_f\omega$ = frequency swing in cycles per second.
μ = modulating-signal frequency in cycles per second.

Methods of producing frequency and phase modulation are discussed more fully in Section 8.

BIBLIOGRAPHY

Crosby, Murray G., Carrier and Side Frequency Relations with Multitone Frequency or Phase Modulation, *RCA Rev.*, Vol. 3, 103–106 (July 1938).
Gray, and Mathews, *A Treatise on Bessel Functions*, Macmillan (1931).
Hund, August, *Frequency Modulation*. McGraw-Hill (1942).
McIlwain, K., and J. G. Brainerd, *High-Frequency Alternating Currents*. John Wiley (1939).
Morecroft, J. H., *Principles of Radio Communication*. John Wiley (1933).
Proc. I.R.E.

DETECTORS

By Vernon D. Landon

Demodulation, or detection, is the process whereby a wave resulting from modulation is so operated upon that a wave is obtained having substantially the characteristics of the original modulating wave.

Square-law detection is that form of detection which depends on the fact that e^2 of the power series for a non-linear circuit excited by a modulated wave contains the difference frequency between the carrier and one of its sidebands, which is the original modulating frequency.

Linear detection is that form of detection in which the output voltage under consideration is substantially proportional to the instantaneous peak carrier voltage throughout the useful range of the detecting device.

18. SQUARE-LAW DETECTION

Any device having a non-linear current-voltage characteristic will serve as a detector. Such devices include certain crystals and thermionic vacuum tubes. (See Figs. 1, 2, and 3.) The applied modulated voltage is of the form given in eq. (5), p. 7-72, which, when squared, gives audio-frequency terms of the form

$$\begin{aligned} i_{pa} = -\frac{r_p}{2}\frac{\partial r_p}{\partial e_p}\Bigg[& \frac{E_sE_n \cos(\omega_n t + \phi_1)}{Z'_s Z'_{(s+n)} Z'_n} + \frac{E_sE_n \cos(\omega_n t + \phi_2)}{Z'_s Z'_{(s-n)} Z'_n} \\ & + \frac{E_sE_m \cos(\omega_m t + \phi_3)}{Z'_s Z'_{(s+m)} Z'_m} + \frac{E_sE_m \cos(\omega_m t + \phi_4)}{Z'_s Z'_{(s-m)} Z'_m} \\ & - \frac{E_nE_m \cos(\omega_n t - \omega_m t + \phi_5)}{Z'_{(s+n)} Z'_{(s+m)} Z'_{(n-m)}} - \frac{E_nE_m \cos(\omega_n t - \omega_m t + \phi_6)}{Z'_{(s-n)} Z'_{(s-m)} Z'_{(n-m)}} \\ & - \frac{E_nE_m \cos(\omega_n t + \omega_m t + \phi_7)}{Z'_{(s-n)} Z'_{(s+m)} Z'_{(n+m)}} - \frac{E_nE_m \cos(\omega_n t + \omega_m t + \phi_8)}{Z'_{(s+n)} Z'_{(s-m)} Z'_{(n+m)}} \\ & - \frac{E_n^2 \cos(2\omega_n t + \phi_9)}{Z'_{(s+n)} Z'_{(s-n)} Z'_{(2n)}} - \frac{E_m^2 \cos(2\omega_m t + \phi_{10})}{Z'_{(s+m)} Z'_{(s-m)} Z'_{(2m)}} \Bigg] \end{aligned} \qquad (1)$$

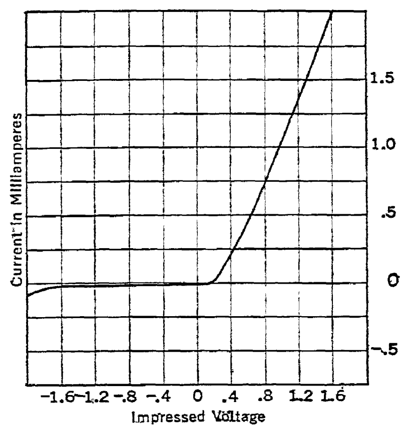


Fig. 1. Static Characteristic of Iron Contact on Ferro-silicon, 25% Fe

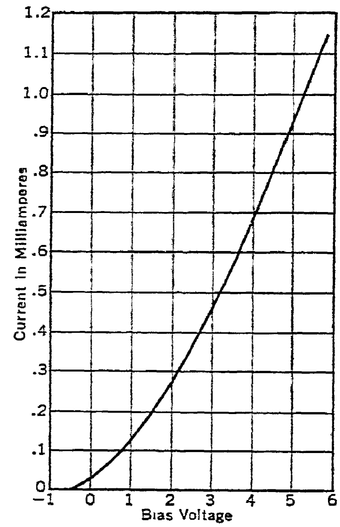


Fig. 2. Static Characteristic of Diode

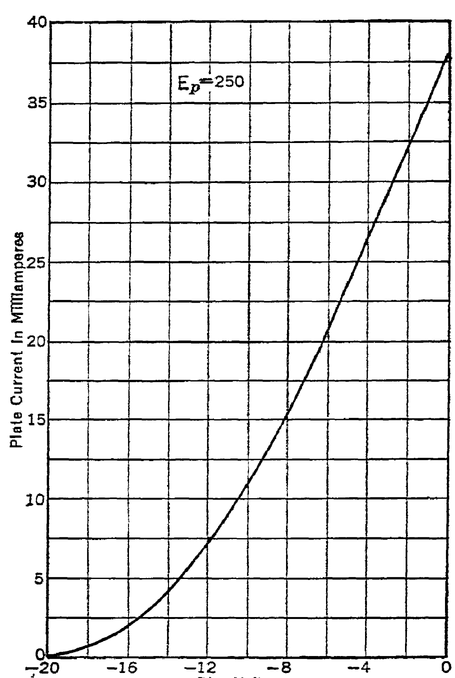


Fig. 3. Static Characteristic of Triode

There are two terms present of each original signal frequency, resulting from the beating of two sidebands against the carrier. From this it is seen that there will not be detection unless the carrier is present, but that the *carrier and one sideband* are sufficient. If regard is paid to differences of phase, terms of the same frequency may be added; the terms present and their phase angles are given in Table 1.

Table 1. Square-law Detection

Simplified Expression for the Audio-frequency Current resulting from Square-law Detection of a Carrier and Both Side Bands. Similar Terms Appear for each Pair of Audio Components.

Term *	Phase Angle Triple-primed Values of Resistance, etc. Apply to Original Modulator at Sending End.
$\dfrac{2\,E_s\,E_n(r''' + r'''_p)\cos(\omega_n t + \beta'_1)}{z'''_s\,z'^2_s\,z'_n}$	$\theta_n - \tan^{-1}\dfrac{x'''_n}{r''' + r'''_p} - \tan^{-1}\dfrac{x_n}{r + r_p}$
$\dfrac{2\,E_s\,E_m(r''' + r'''_p)\cos(\omega_m t + \beta'_2)}{z'''_s\,z'^2_s\,z'_m}$	$\theta_m - \tan^{-1}\dfrac{x'''_m}{r''' + r'''_p} - \tan^{-1}\dfrac{x_m}{r + r_p}$
$-\dfrac{2\,E_n\,E_m\cos(\omega_n t - \omega_m t + \beta'_2)}{z'^2_s\,z'_{n-m}}$	$\theta_n - \theta_m - \tan^{-1}\dfrac{x'''_n}{r''' + r'''_p} + \tan^{-1}\dfrac{x'''_m}{r''' + r'''_p} - \tan^{-1}\dfrac{x_{n-m}}{r + r_p}$
$-\dfrac{2\,E_n\,E_m\cos(\omega_n t + \omega_m t + \beta'_4)}{z'^2_s\,z'_{n+m}}$	$\theta_n + \theta_m - \tan^{-1}\dfrac{x'''_n}{r''' + r'''_p} - \tan^{-1}\dfrac{x'''_m}{r''' + r'''_p} - \tan^{-1}\dfrac{x_{n+m}}{r + r_p}$
$-\dfrac{E_n^2\cos(2\,\omega_n t + \beta'_5)}{z'^2_s\,z'_{2n}}$	$2\theta_n - 2\tan^{-1}\dfrac{x'''_n}{r''' + r'''_p} - \tan^{-1}\dfrac{x_{2n}}{r + r_p}$
$-\dfrac{E_m^2\cos(2\,\omega_m t + \beta'_6)}{z'^2_s\,z'_{2m}}$	$2\theta_m - 2\tan^{-1}\dfrac{x'''_m}{r''' + r'''_p} - \tan^{-1}\dfrac{x_{2m}}{r + r_p}$

* Each term must be multiplied by $-\dfrac{r_p}{2!}\dfrac{\partial r_p}{\partial e_p}$.

With only one audio component present the per cent modulation is $200E_n/E_s = M$. Hence the per cent second harmonic present, in terms of the fundamental, is $M/4$. For large per cent modulation this system is therefore not satisfactory. When more than one signal frequency is present as modulation on the carrier wave, extraneous frequencies are produced corresponding to the sum and difference frequencies of all the signal components taken in pairs.

The square-law detector is now largely used only for detection of carriers having small percentage modulation and for the reception of single sideband signals. It is the only form which will work with these latter. In detecting single sideband and carrier, second harmonics of the audio components are not produced, but frequencies corresponding to the sums and differences between the frequencies of the components of the original signal are present.

SQUARE-LAW DETECTOR CIRCUITS. Any non-linear circuit excited by a voltage small enough so that the static characteristic over the region used is closely a parabola is satisfactory as a square-law detector. When the vacuum tube is used, detection may

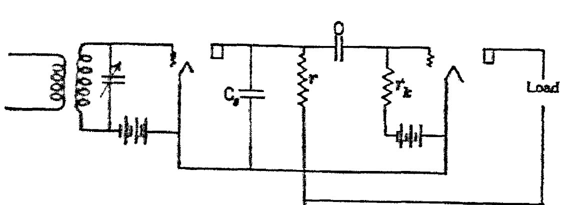


Fig. 4. Plate Current Detector, Resistance Coupled to Succeeding Amplifier Tube

occur either in the grid or plate circuits. Figure 4 shows a circuit which is biased practically to plate-current cutoff so that detection takes place owing to the curved mutual characteristic.

The voltage at any frequency k impressed on the amplifier tube is

$$E_{LK} = \frac{A_K \mu^2 r_p z_K}{z'^2_s z'_K} \frac{\partial r_p}{\partial e_p} \tag{2}$$

For high detected voltage the external plate impedance at the carrier frequency (z_s) should be small, but that at the signal frequencies (z_k) should be large. The high audio impedance is shunted by C_s to achieve this result.

Figure 5 shows a circuit in which detection occurs in the grid. The grid and cathode elements act as a diode detector. The a-f currents, resulting from demodulation, produce an a-f voltage across the grid leak and so between grid and cathode. This voltage controls the electron stream so the tube also acts as an amplifier. A high audio impedance and low radio impedance are necessary so that the grid leak is shunted by a small condenser.

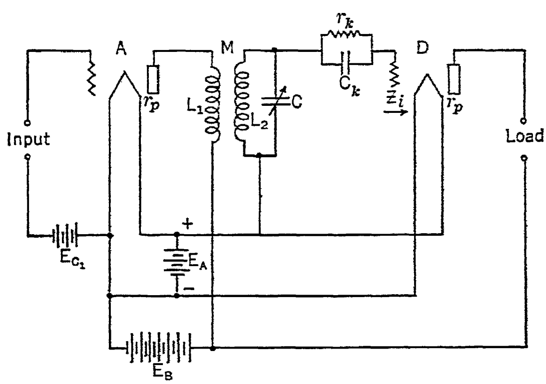


FIG. 5. Grid Current Detector (D), Transformer Coupled to Preceding Tube (A)

This type of detector broadens the tuning of the preceding tuned circuit (owing to the low impedance of the grid when positive), and overloads badly on strong signals, owing to the r-f and d-c components of the plate current. It is very sensitive, however.

19. LINEAR DETECTION

When very large signals are applied across non-linear circuits the power series convergence based on the static characteristics is so slow that the method is not useful. The equation for the modulated wave can optionally be written in the form (see eq. 6, p. 7-72):

$$e = E_s \left[1 - 2 \sum \frac{E_K}{E_s} \cos(\omega_K t + \theta_K) \right] \cos(\omega_s t + \theta_s) \tag{3}$$

which represents a voltage, of fixed frequency, whose amplitude is slowly varying about a mean value E_s. The sum (Σ) within the brackets will represent a voltage of the same wave shape as the original modulating voltage, provided the modulation and transmission have been distortionless. The instantaneous peak voltage of the carrier current is

$$E_m = E_s \left[1 - 2 \sum_K \frac{E_K}{E_s} \cos(\omega_K t + \theta'_K) \right] \tag{4}$$

Rectification diagrams of the detecting device are obtained by plotting the total *direct current* in the detector circuit against various *steady* values of rms (or peak) carrier voltage. (See Fig. 6.) Then when E_m is varied (in accordance with the signal voltage) this current varies; this detected current is equivalent to a steady direct current, whose value is determined by the value of direct current which flows when $E_m = E_s$, on which is superposed a varying current which will follow the variations in E_m (and so in the signal voltage) more or less closely, depending on the linearity of the rectification diagram. Since the periodicity (ω_s) of the carrier voltage is not of importance except in fixing the circuit impedances, the rectification diagrams can be obtained by using any convenient frequency, such as 60 cycles, rather than an actual carrier frequency, provided the circuit impedances are made equal in the two cases.

It will be noticed from the rectification diagrams shown in Figs. 7–12 that the increment in direct current due to a varying grid voltage varies approximately as the square of the maximum value of this voltage when E_m is small (as required by the power series expansion), but that, as E_m increases, the d-c response varies *linearly* with E_m. When it does so the distortion introduced in detection will be small.

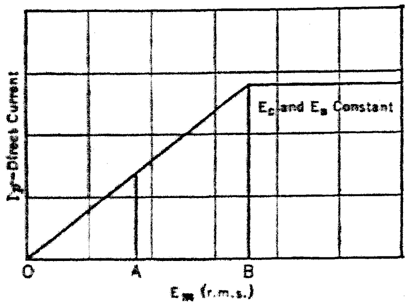


FIG. 6. Ideal Detector Characteristic (Rectification Diagram)

As an illustration, consider the plate-current detector shown in Fig. 4. Currents of radio frequency are by-passed by C_s; hence the only current which flows through r when E_m is steady is a direct current. If the value of this current were varied in accordance with the signal voltage, the varying voltage impressed on the grid of the succeeding amplifier tube would be simply the signal voltage. If the variation of direct current against E_m were given by the curve of Fig. 6 and the average value of E_m were equal to OA, then the variation of current would be exactly proportional to the signal voltage, *even for 100 per cent modulation*, provided $E_m < OB$. This means that there would be *no distortion* introduced by the detector.

Although no device available at present has exactly the characteristic of Fig. 6, certain crystals approximate it, and the vacuum tube can be made to approach it by a proper

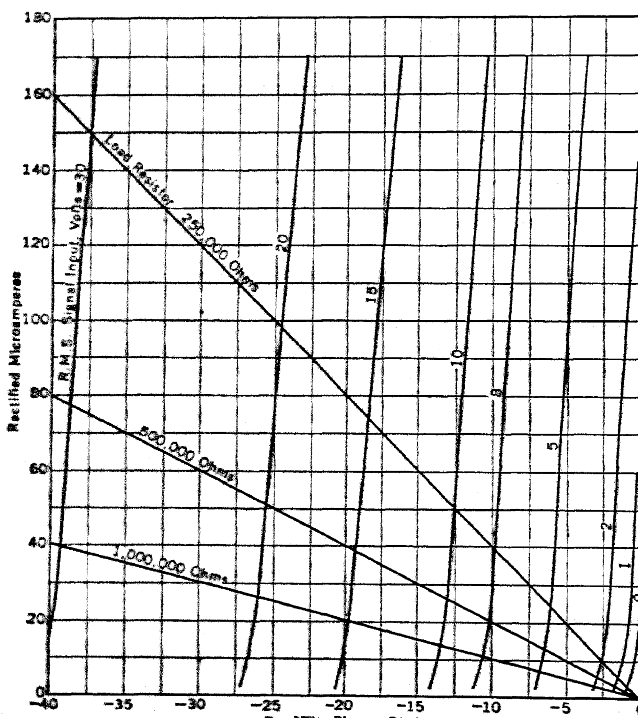


FIG. 7. Rectification Diagram for Diode Detector

choice of operating conditions and load. Figure 7 shows a set of curves for a diode from which rectification diagrams for any load resistance can be obtained; the diagram for 500,000-ohm load is shown in Fig. 8, which could also have been obtained experimentally. A diode gives the nearest approach to linear detection of any known detector.

When the characteristic is not a straight line, a Taylor's series expansion can be used in terms of e_{pa} and the change (e_m) in E_m:

$$i_{pa} = e_{pa}\frac{\partial i_p}{\partial e_p} + e_m\frac{\partial i_p}{\partial e_m} + \frac{1}{2}\left\{e_{pa}^2\frac{\partial^2 i_p}{\partial e_p^2} + 2e_{pa}e_m\frac{\partial^2 i_p}{\partial e_p \partial e_m} + e_m^2\frac{\partial^2 i_p}{\partial e_m^2}\right\} + \quad (5)$$

where i_{pa} is the current due to the variations e_{pa} and e_m, i_p is the plate-current ordinate of the rectification diagram, and e_{pa} is the plate potential due to i_{pa} and e_m. The reciprocal of $\partial i_p/\partial e_p$ is the analog of the plate resistance used in Section 5, and has been named the *detection plate resistance*. The derivative $\partial i_p/\partial e_m$ is the slope of the rectification curve—it is analogous to the mutual conductance of a triode and is called the *detection mutual conductance* or the *transrectification factor*.

All the calculations given for class A amplifiers now may be utilized in the detector circuit, including calculations of distortion, detection efficiency, etc.

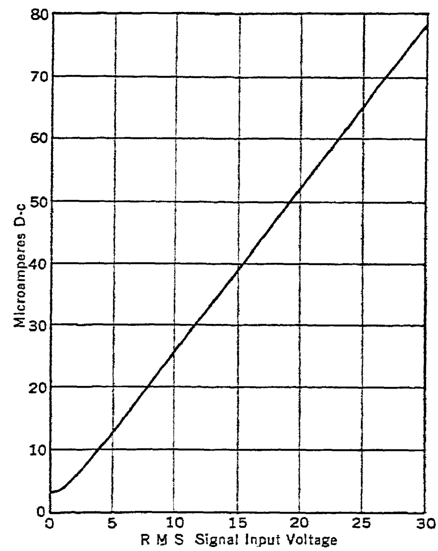


FIG. 8. Load Rectification Diagram for Diode Using 500,000-ohm Load

LINEAR DETECTORS. As stated above, the diode gives the nearest approach to linear detection known today. It has the objection of putting a fairly small resistance in parallel with a tuned circuit and so considerably broadening the tuning. When a diode is used the applied r-f voltage should be made as high as the tube can stand and a load resistance selected to give minimum distortion.

The triode is not as distortionless as the diode but sometimes fits in better with the circuit requirements. Rectification diagrams of a triode alone and with load are given in Figs. 9 and 10. In general, the plate voltage should be as high as possible, limited only by the supply available or the limits of the tube itself (flashover or heating).

When the plate voltage has been determined, rectification diagrams of the tube alone at various grid biases are obtained. If it is desirable, as is usual, to use the tube with carrier voltages which are 100 per cent modulated, the plate current should not reduce to zero for any finite value of E_m, or part of the modulation will produce no effect. In other words, the tube should not be biased below cutoff. The best results will be obtained if the tube is biased at, or slightly above, cutoff (on the static characteristic); this condition will be indicated on the rectification diagram by a small plate current for $E_m = 0$. For the tube of Fig. 9 the grid voltage is thus determined as either -1.5 or -3.0 volts.

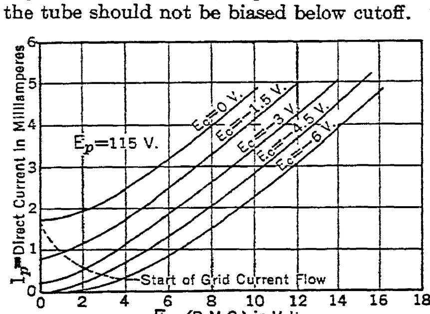


FIG. 9. Rectification Diagrams of Triode (Alone). All data were taken with 60-cycle impressed voltage.

To determine the proper load resistance, rectification diagrams of the tube with different loads (see Fig. 10) are obtained. If E_m is varied over a straight portion of one of these curves, distortion-free detection will result; if the variation is over a curved portion the distortion may be calculated as for an amplifier. (See p. 7-07.) For the tube used here a load impedance of 30,000 ohms would cause the tube to introduce a very low percentage of distortion for a carrier having an average rms value of 6.25 volts, and 100 per cent modulated. The difference between curves C and E illustrates the gain in efficiency, but increased distortion, obtained by by-passing the load resistance.

When the triode is used as a detector with large signal voltages of high percentage of modulation, operation with minimum distortion is always accompanied by a positive grid swing and so a flow of grid current. This causes a low input impedance and so decreases the selectivity of the tuned circuit just preceding the detector.

The use of a tetrode (see Fig. 11) makes it possible to obtain rectification diagrams which reach saturation for negative values of control grid voltage, so giving straighter operating curves without lowering the input impedance of the tube. The same method of analysis is followed, the only additional factor being the proper voltage to use for the screen grid. The method employed in selecting the screen-grid voltage depends to a large extent upon the type of plate load. In general, a higher screen-grid voltage gives a better rectification diagram (straighter and steeper). However, when there is a large d-c resistance in the plate circuit, the plate voltage falls considerably for large signal voltages, so that, if the screen-grid voltage is comparable with the plate voltage, electrons will land on the screen grid rather than on the plate.

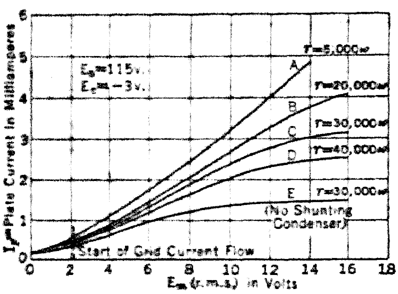


FIG. 10. Load Rectification Diagrams of Triode. For all curves but E the load was shunted by a 2 μf condenser. All data were taken with 60-cycle impressed voltage.

When reactance coupling is used the screen-grid voltage should be the maximum consistent with the plate voltage. For this type of coupling the d-c resistance of the detector plate load is small enough so that large carrier voltages will not cause sufficient drop for the plate voltage to become comparable to the screen-grid voltage. Therefore, there is no necessity of keeping the screen voltage low to prevent saturation (on the rectification diagram), and the gain in mutual conductance with high screen voltage raises the sensitivity and maximum output as well as decreases the curvature of the rectification characteristic.

When resistance coupling is used the selection of the optimum screen-grid voltage is somewhat more difficult. The problem is to select a value of screen-grid voltage that will give reasonable detector output at the normal values of input and yet not cause overloading for the larger values of input voltage. Two rectification diagrams for different values of screen-grid voltage and a high resistance load are shown in Fig. 12. The curve for

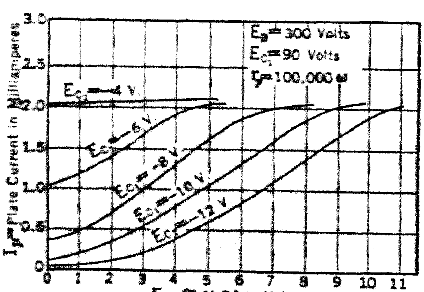


FIG. 11. Load Rectification of a Tetrode. All measurements were made at 60 cycles.

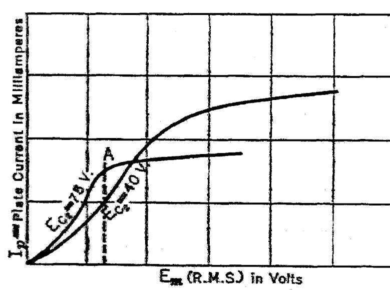


FIG. 12. Load Rectification Diagrams of a Tetrode, Showing the Effect of Screen Grid Voltage on Saturation. (Resistance load.)

E_{c_2} = 75 volts is better than that for E_{c_2} = 40 volts for inputs up to a point A, but the tube overloads for voltages greater than A. It is thus necessary to know the maximum signal voltage which will be impressed on the detector, and the screen-grid voltage which will just not overload at that voltage should be chosen.

BIBLIOGRAPHY

Bell Sys. Tech. J.
Exper. Wireless and Wireless Engineering.
J. A.I.E.E.
Proc. I.R.E.
Ballantine, S., Detection at High Signal Voltages—Plate Rectification with the High-vacuum Triode, *Proc. I.R.E.*, Vol. 17, 1153 (July 1929).
Everitt, W. L., *Communication Engineering.* McGraw-Hill (1937).
McIlwain, Knox, and J. G. Brainerd, *High-frequency Alternating Currents.* Wiley (1939).
Pierce, G. W., *Electric Oscillations and Electric Waves.* McGraw-Hill (1920).
Terman, F. E., *Radio Engineering.* McGraw-Hill (1937).

OSCILLATORS

By Carl C. Chambers

An oscillator is a non-rotating device for producing alternating current, the output frequency of which is determined by the characteristics of the device. Oscillators can be conveniently divided into three classifications: (1) vacuum-tube oscillators, (2) electromechanically controlled oscillators, and (3) spark-gap oscillators.

20. VACUUM-TUBE OSCILLATORS

The most generally used oscillator in the field of communication engineering is the vacuum-tube oscillator. Since a vacuum tube can act as an amplifier of power, it can act as an oscillator. The power needed in the grid circuit can be supplied by the plate circuit. A vacuum-tube oscillator circuit acts as a power converter, changing d-c power into a-c power, having a frequency determined by the parameters of the circuit. The efficiency of vacuum-tube oscillators can be above 90 per cent, although for many purposes they operate at low efficiency. Vacuum tubes are constructed to dissipate as high as 125,000 watts, the power output of oscillators using such tubes being much greater.

SIMPLE OSCILLATOR CIRCUITS. In ordinary vacuum-tube oscillators, power is fed into the grid circuit from the plate circuit by means of either electrostatic or electromagnetic coupling between these circuits. When sufficient voltage of proper phase is introduced into the grid circuit, the a-c component of the plate current will persist, or, in other words, the circuit will oscillate.

Many oscillator circuits have been devised, several of which are shown in Fig. 1. In each of these circuits, the oscillatory circuit is the mesh containing L and C. In the Hartley circuit, the inductance L is tapped (not necessarily at the center); in the Colpitts circuit, the capacitance C is the series capacitance of C_1 and C_2. The tuned-plate tuned-grid oscillator has two separate oscillatory circuits, one in the plate circuit and one in the grid circuit, both tuned to approximately the same frequency. In each of these oscillators the load is coupled to the oscillatory circuit. This coupling is usually but not always inductive coupling.

The voltage introduced into the grid circuit due to the plate current, in (*a*) and (*b*), is the drop across part of the oscillatory circuit. In (*c*), (*e*), and (*f*), it is the voltage across one winding of the transformer, one winding of which is the inductance, L, of the oscillatory circuit. And in (*d*) it is the voltage drop across the grid oscillatory circuit due to the current through it and the grid to plate capacitance. The peak value of this feedback voltage in each case must be great enough to cause oscillation to persist. For stable operation, it should be roughly two and one-half to three and one-half times the d-c grid voltage necessary to bias the tube to plate current cutoff at the operating d-c plate potential.

The capacitances, C_B, are by-pass condensers, and the inductances, L_B, are choke coils arranged to prevent the alternating current from passing through the d-c supply source, E_B, or through the biasing resistor or grid leak, r_B. The grid voltage in most oscillators goes positive for part of the cycle. Thus the grid bias voltage is conveniently obtained by means of a grid leak and condenser. This method of biasing the oscillator tube tends to self-adjust the grid bias to permit operation over a wide range of values of the feedback voltage, since with greater excitation the grid tends to go more positive and consequently the grid bias becomes greater, resulting in essentially the same plate-current flow. Some oscillators have a comparatively small external grid bias (or self-bias by means of a cathode resistor) in addition to the grid leak. This externally supplied bias is a safety device to prevent too much plate current from flowing if the oscillations should cease for any reason.

NON-LINEAR THEORY OF OSCILLATIONS. The complete solution of even the simplest oscillator circuits is extremely complex. However, van der Pol, by means of very radical assumptions, has obtained the solution of several oscillator circuits. One of these solutions is for the circuit shown in Fig. 1*f*. The following assumptions are made concerning the circuit: (1) that the grid bias voltage is obtained by means of an external electromotive force; (2) that no current flows in the grid circuit; (3) that the capacitances C_B are so large that their a-c impedance is zero; and (4) that μ, the amplification factor, is constant. These approximations make the solution useful only from a qualitative standpoint.

The necessary condition for oscillation to persist is found to be that α, the damping constant of the circuit taken as a whole, is negative. To a rough approximation the amplitude of the oscillation is proportional to the square root of the ratio of α to the coefficient of the cube term in the power series expansion of the characteristic of the vacuum tube.

The wave form of the oscillation is dependent upon ϵ, the product of α and LC. The solutions of the circuit for several values of ϵ as calculated on the differential analyzer at

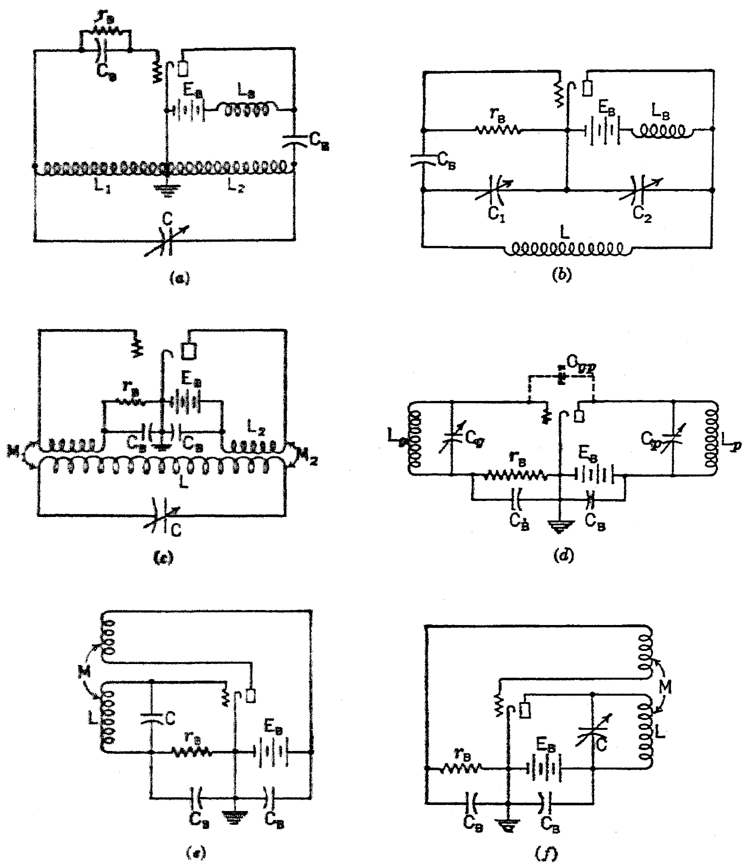


FIG. 1. Conventional Oscillator Circuits

the Moore School of Electrical Engineering are shown in Fig. 2. For small values of ϵ, the oscillations are essentially sinusoidal, the frequency being given by $\dfrac{1}{2\pi\sqrt{LC}}$. For large values of ϵ, the oscillations depart radically from sine waves. This form of oscillation has been called *relaxation oscillation*. It occurs in the multivibrator, the human heart, the hydraulic ram, etc. Van der Pol gives the period for $\epsilon \gg 1$ as roughly equal to 2ϵ.

CONDITIONS FOR SELF-OSCILLATION. For small amplitudes of oscillation, the non-linear resistances of the plate and grid circuits can be approximated by linear resistances. The condition for these small oscillations to build up is that the damping constant is negative. This is the condition for unstable systems in transient circuit theory. However, as the oscillations increase in amplitude, the linear circuit theory fails.

Llewellyn has shown that, under steady-state operation for any frequency component of the current, a non-linear resistance can be replaced by an equivalent impedance which can usually be roughly approximated by a pure resistance. When the circuit oscillates normally, the fundamental frequency of oscillation has a constant amplitude so that the plate to filament circuit in the tube can be replaced by such an effective resistance. In order for steady-state currents of a given frequency to flow in a circuit containing no emf of that frequency, the determinant of the coefficients of the currents in the mesh equations must be zero. An oscillator must satisfy such a condition. The approximate resulting conditions are noted for most of the circuits in Fig. 1. The value of the effective plate resistance, r_p, for any amplitude is given approximately by the inverse slope of the secant joining the limiting points of operation on the plate characteristic of the tube. By the reverse calculation the magnitude of oscillation for any given circuit can be predicted.

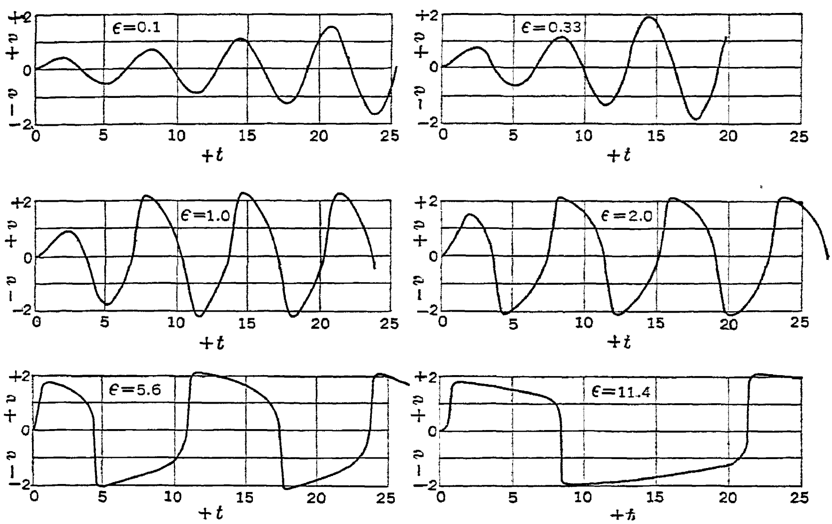


FIG. 2. Solutions of van der Pol's Equation for the Non-linear Theory of Oscillations:

$$\left(\frac{d^2v}{dt^2} - \epsilon(1 - v^2)\frac{dv}{dt} + v = 0\right)$$

CURRENT AND VOLTAGE RELATIONS IN SIMPLE OSCILLATOR CIRCUITS. As stated at the beginning of this chapter, the vacuum-tube oscillator is simply an amplifier arranged so that the power needed in the grid circuit is supplied by the plate circuit. Ordinarily, the amplifier so arranged belongs to that group known as class C amplifiers.

The current and voltage relations of a typical oscillator are shown in Fig. 3 for one complete cycle. The plate current flows during the portion of the cycle when the instantaneous plate voltage is least so that, since the energy dissipated at the plate of the vacuum tube is the integral of the instantaneous plate voltage times the instantaneous plate current, the tube plate loss is least when the current flows for as small a portion of the cycle as possible. When plate current flows, the grid voltage is positive with respect to the grid bias necessary for plate-current cutoff. Thus the duration of plate current decreases and consequently the plate efficiency increases with simultaneous increases in the grid bias and the a-c grid voltage. However, increasing the grid voltages in this way increases the losses in the grid circuit. The a-c grid voltage usually has a peak value of two and one-half to three and one-half times the grid bias necessary for plate current cutoff at the d-c plate voltage.

Since plate current flows only during minimum instantaneous plate voltage, the minimum plate voltage is of major importance. The impedance of the oscillatory circuit at resonance is adjusted so that, for the normal power output, the peak a-c voltage across it is slightly smaller than the applied d-c plate voltage. Therefore, the plate voltage causing current to flow is small compared with the d-c plate voltage, since it is essentially the d-c plate voltage minus the peak a-c voltage across the oscillatory circuit. Thus the a-c and d-c voltages remain practically equal as the d-c voltage is varied. It is this equality between the d-c and a-c voltages that makes the plate modulation of oscillators and class C amplifiers so nearly linear.

In ordinary triodes, the maximum plate current flows for a positive value of the grid voltage somewhat less than the minimum instantaneous plate voltage. In most oscillators at full load, the relation between the plate voltage and the grid voltage is such that the maximum instantaneous grid voltage and the minimum instantaneous plate voltage are respectively slightly greater and less than voltages satisfying the condition of maximum plate current.

When the bias is obtained by means of a grid leak and condenser, the a-c voltage introduced into the grid circuit has only a negligible effect on the maximum positive grid voltage, this being controlled by the magnitude of the resistance of the grid leak. The value of the grid leak is best determined by trial and error until the peak positive grid voltage is about 0.8 of the minimum instantaneous plate voltage. In small oscillators

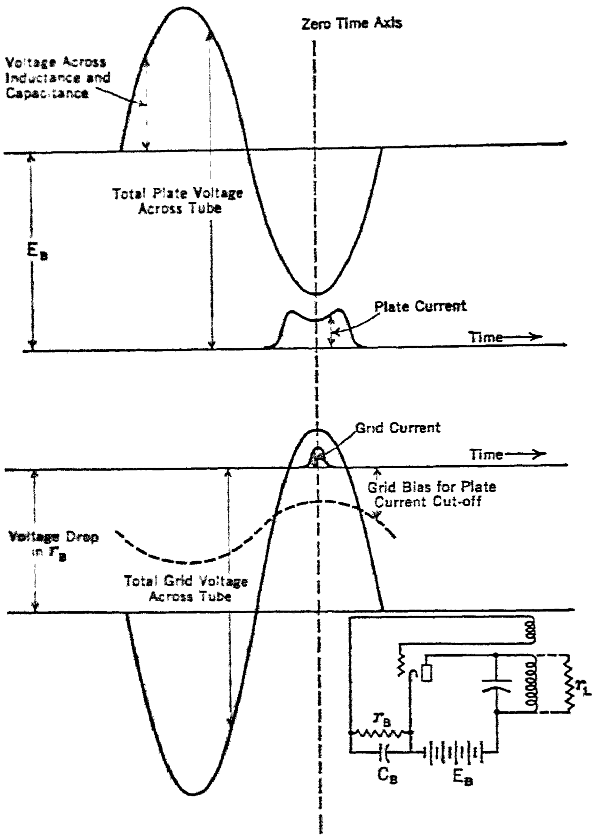


Fig. 3. Current and Voltage Relations of a Representative Oscillator

where efficiency is not vital the value of the grid leak can vary over wide limits; a value anywhere from a few thousand ohms to a megohm may be determined by trial and error until the oscillator operates satisfactorily.

Sometimes when the time constant of the grid leak and condenser is high and when the ratio of the a-c plate voltage to the a-c grid voltage is low, "blocking" will take place. Blocking is a relaxation oscillation and is due to the grid being at a higher potential than the plate so that the secondary emission from the grid causes the grid leak and condenser to develop a high bias which stops the oscillation until the charge leaks off the condenser, at which time oscillations again start and the cycle repeats itself. It can be corrected by increasing the ratio of the a-c plate voltage to the a-c grid voltage.

OSCILLATORY CIRCUIT DESIGN. The heart of any sinusoidal oscillator is the oscillatory circuit consisting of an inductance and a capacitance connected in series. Energy is supplied to this circuit from a d-c source by means of a vacuum tube, and energy is taken away from it by coupling the load to it in one of various ways. The circuit performs a function similar to that of the flywheel on a reciprocating steam engine.

The oscillatory circuit must store the energy used by the load long enough so that it can supply the load continuously even though it receives energy only during a small portion of the cycle. It has been found empirically that in order to do this job effectively, i.e., to give a good wave form, the ratio of the peak energy stored per cycle must be at least twice the energy fed into the load per cycle. From ordinary circuit theory it is known that a load resistance paralleling a tuned circuit, or coupled to a tuned circuit, can be replaced approximately by a resistance in series with the tuned mesh. When this is done the ratio of the peak stored energy to the energy dissipated per cycle is

$$\frac{V^2Cf}{I^2r} = \frac{VI}{2\pi I^2r} = \frac{fL}{r} \tag{1}$$

where V is the rms voltage across the capacitance, C, and across the inductance, L; I is the rms current through the inductance and capacitance; $r(= L/r_LC$, where r_L is the shunting resistance load) is the effective series resistance of the circuit, assumed to be small compared with 2π times the frequency, f, times L. Thus if the ratio is to be greater than 2, $2\pi fL/r(= Q)$ must be greater than 4π. This Q refers to the inductor together with the equivalent resistance due to the load.

On the other hand, if the ratio $L/r(= r_LC = r_L/4\pi^2f^2L)$ is made too large the resonant resistance of the circuit will be too high to obtain a satisfactory power output. It is therefore necessary to compromise between power output and wave form. It is in general good practice to make the ratio of the stored power to the power output about 2, i.e., to make $Q = 4\pi$. L and C are then given in terms of the power, the frequency, and the voltage (essentially the d-c plate voltage) as

$$L = \frac{V^2}{8\pi^2Pf} \tag{2}$$

$$C = \frac{2P}{V^2f} \tag{3}$$

These formulas are not meant to be critical. If good wave form is more important, use a value of P in these formulas greater than the actual power; if power output is more important use a value of P less than the actual power. This principle is frequently stated by saying that the wave form is improved by decreasing the L to C ratio of a tuned circuit. These circuit considerations apply equally well to the tuned circuit of a class C amplifier.

CONSTANT-FREQUENCY OSCILLATORS. Using the equivalent impedance discussed under Conditions for Self-oscillation, above, (*a*) and (*b*) of Fig. 1 are special cases of the equivalent circuit shown in Fig. 4. The ordinary mesh equations for this circuit are

$$I_p(z_p + z_1 + z_5) - I_1(z_1 + z_m) + I_g(z_m + \mu z_g) = 0 \tag{4a}$$

$$-I_p(z_1 + z_m) + I_1(z_1 + z_2 + z_3 + 2z_m) - I_g(z_2 + z_m) = 0 \tag{4b}$$

$$I_pz_m - I_1(z_2 + z_m) + I_g(z_g + z_2 + z_4) = 0 \tag{4c}$$

Most of these impedances are functions of the frequency. And, since the condition that any of these currents can be other than zero is that the determinant of the coefficients of the I's is zero, the frequency of oscillation is the frequency which will make that determinant zero. Thus the frequency depends not only upon the parameters of the oscillatory circuit but also on the other parameters of the circuit. In order that the frequency shall remain constant, all these parameters must in general remain constant.

To keep the inductances, capacitances, and resistances external to the tube constant is a problem in the design and temperature control of those parts. On the other hand, regardless of the design, the effective impedance of the tube itself changes with use and supply voltages. Several methods have been devised to maintain essentially constant supply voltages, and seasoned tubes tend to reduce the changes in the tubes themselves. However, by adjusting the other circuit parameters the dependence of the frequency on the tube impedances can be minimized.

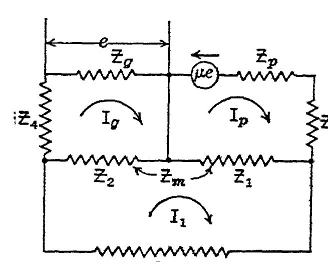


FIG. 4. The Equivalent Circuit of Most of the Circuits of Fig. 1

The most common method is simply to increase the sharpness of the oscillatory circuit, that is, decrease the decrement of the circuit. This is carried to the extreme by the use of mechanical resonators such as the quartz crystal and the magnetostriction rod.

Llewellyn has discovered that, by proper adjustment of the impedances Z_4 and Z_5, it is possible in the limiting case of no harmonic currents in the tube to eliminate the dependence of the frequency of oscillation on the tube impedances. Although this limiting case is never reached, the use of these impedances tends to stabilize the frequency. In addition it is sometimes possible to change the frequency by a simple change in the parameters of the oscillatory circuit without disrupting this independence of the frequency on the tube impedances. A circuit derived from the Colpitts circuit for this purpose is shown in Fig. 5, where $L_4 = L_3C_1/C_2$. Thus the frequency can be adjusted by changing C_1 and C_2 proportionally without changing the calculated value of L_4.

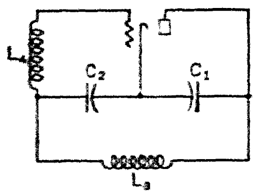


FIG. 5. A Constant Frequency Circuit Derived from a Colpitts Oscillator

THE HARMONIC CONTENT OF OSCILLATORS. From the non-linear theory of oscillators (see article 20) it is known that an oscillator is stable only if it is non-linear. Thus an oscillator always produces harmonics in its output. The percentage of the harmonics becomes small when ϵ becomes small, that is, when the feedback from the plate circuit to the grid circuit is small. A circuit applicable at low frequencies in which the feedback can be controlled is shown in Fig. 6. The variable resistance r controls the amount of feedback. This circuit not only has a controlled harmonic content, but it also is very stable with frequency. It is excellent for a laboratory oscillator where r can always be adjusted to compensate for changes in the tube characteristic.

In large oscillators and class C amplifiers a series tuned circuit is sometimes connected in series with the plate circuit to offer a high impedance to the harmonics. Such a tuned circuit, in addition to decreasing the harmonic content to a marked extent, increases the efficiency.

SEPARATELY EXCITED OSCILLATORS. Separately excited oscillators are a special group of class C amplifiers. The voltage and current relations of the true oscillator are essentially those of a properly operated class C amplifier. If, however, the frequency of the exciting voltage differs from that at which the tube would oscillate, the plate-current wave form is not symmetrical about the zero ordinate shown in Fig. 3. Increasing the dissymmetry of this wave by varying the capacitance of the oscillatory circuit increases the average or d-c current. Thus the frequency of the oscillatory circuit is adjusted in this type of class C amplifiers until the d-c current is a minimum.

Most of the theory and empirical relations given for oscillators can be applied to class C amplifiers.

SYNCHRONIZATION OF OSCILLATORS. When two oscillators oscillating at neighboring frequencies are loosely coupled together, they mutually distort the voltage drop across their oscillatory circuits. This distortion in turn distorts the current and voltage relations discussed above. This distortion causes a magnified distortion of the plate current since even a small per cent change in the oscillatory circuit voltage has a great effect on the minimum plate voltage. This change in plate current causes an additional change in the voltage by affecting the oscillatory circuit drop directly and indirectly through the change in grid excitation. Thus the voltage due to the one oscillator causes a magnified change in the other oscillator. This distortion tends to shift the phase of the plate voltage with respect to the grid voltage. When this phase shift exceeds the limit allowable for stable oscillation, the oscillator jumps into oscillation at the frequency of the other oscillator.

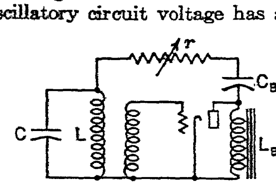


FIG. 6. An Oscillator for Low Harmonic Distortion

The vulnerability of an oscillator to synchronization increases with the L to C ratio of the oscillatory circuit. Thus, when two oscillators are to oscillate independently even though there is a coupling between them, the L to C ratio should be made as small as is consistent with stable oscillation.

In order for one oscillator to oscillate at the frequency of a master oscillator without changing the frequency of the master oscillator, it is necessary that the coupling between the oscillators be unidirectional. That is, the reaction of the secondary oscillator on the master oscillator must be negligible. Such coupling can be obtained by means of the ordinary amplifier. Using several such amplifiers, each fed by the master oscillator, the frequency of several oscillators can be controlled by the frequency of the master oscillator.

An oscillator can be synchronized at a subharmonic of an introduced voltage. The tendency to synchronize with a subharmonic of the introduced voltage is not as great as at the fundamental, and oscillations so obtained are relatively unstable.

BEAT FREQUENCY OSCILLATORS. In order to satisfy the conditions for stable oscillation many of the circuit parameters must be changed when the frequency is changed over a relatively large range. For continuously variable frequency over a wide range, especially over the audio range, it is difficult to make the necessary changes in these parameters. For the same absolute change in frequency beginning at some high frequency, only one parameter, usually the capacitance, need be changed. If the outputs of two oscillators of relatively high frequency are introduced into a square-law detector, and if all the high frequencies are filtered out, only the difference frequency is left. Then as the frequency of one of these oscillators is varied from the frequency of the other to that frequency plus 10,000 cycles, the output of the detector varies in frequency from 0 to 10,000 cycles. Such an arrangement is called a beat frequency oscillator and when properly built is an excellent laboratory instrument.

Care must be taken to insure that the intercoupling between the oscillators is so small that there is no tendency for them to synchronize, since this will introduce distortion in the output of the detector due to the distortion in the oscillators themselves. This can be prevented by mechanical or electromagnetic segregation and balanced bridge circuit feed to the detector, or by amplifiers between the oscillators and the detector.

DYNATRON OSCILLATORS. If, in a vacuum tube, the grid voltage is made more positive than the plate voltage, some of the electrons which attain a high velocity between the cathode and the grid pass through the grid openings, their velocity carrying them on to the plate. These electrons may then knock electrons from the plate. This process is called secondary emission. In some cases secondary emission may exist to such an extent that an increase in plate voltage actually causes a decrease in plate current. Thus, when a tube is operated under these conditions, the a-c plate resistance is negative. A tube operating in this way is called a dynatron.

An oscillatory circuit connected across this negative resistance (Fig. 7) will oscillate provided the absolute value of the negative resistance is less than L/rC. Owing to the small range of plate voltage over which the resistance is negative, these oscillators have not been made to give very large power.

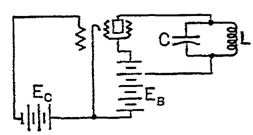


FIG. 7. The Circuit of a Dynatron Oscillator

OSCILLATORS AT HIGH FREQUENCIES. At high frequencies, of the order of 50 Mc, in the case of ordinary receiver tubes, where the time of transit of the electrons between the tube electrodes becomes an appreciable part of the period of the wave, the grid impedance can no longer be considered an open circuit or even a pure capacitor. Under these conditions, the displacement currents arising from the motion of the electrons in the space between the grid and the other electrodes causes a true dissipation of energy in the grid circuit within the tube and a direct conductive coupling to other electrodes. This causes a grid-circuit loading of the tank circuit, decreasing the oscillator efficiency. As the transit-time effect first becomes important, the resistance of the grid due to this decreases approximately as the square of the frequency. Consequently, as the frequency of oscillation is increased through this frequency range, the efficiency decreases rapidly to the point where oscillations can no longer be sustained, even though no external load is coupled to the tank circuit. Decreasing the dimensions and spacing of the electrodes increases the frequency at which transit-time effects become important but correspondingly decreases the power-dissipating capacity of the tube.

At high frequencies, another effect becomes important. The leads within the tubes and the socket are found to have an appreciable impedance. Such impedances must, of course, be considered part of the circuit when that circuit is designed. In many tube types the impedances associated with the leads are so prominent at high frequencies that the highest frequency at which these tubes will oscillate is limited by this consideration rather than by the transit-time effect discussed above.

Since the efficiency of ordinary oscillatory circuits decreases as the frequency for which these circuits are designed increases, it is common to replace coil and condenser type tank circuits by resonant concentric lines. Thus in the tuned-plate tuned-grid oscillator circuit of Fig. 1*d*, two concentric lines are used, one replacing the tuned-plate circuit and one replacing the tuned-grid circuit. The adjustment of the resonant frequency is accomplished by movable short-circuiting plugs in the lines. Various configurations of the concentric lines can be readily adapted to this circuit.

To accomplish the combination of improvements to overcome the adverse effects of transit time and lead impedance, and to adapt the tube for use with concentric line circuit elements, the so-called "lighthouse" tube was developed (see Section 4). Using such tubes, triode oscillators have been operated above 3000 Mc.

THE MULTIVIBRATOR. For large values of ϵ the non-linear theory of vacuum-tube oscillators discussed above indicates that voltages having a discontinuous wave form would be produced. At low frequencies such oscillations can be obtained from a circuit shown in Fig. 1e or f by reducing the capacitance C to zero. The only capacitance is then the distributed capacitance plus the tube capacitance. At low frequencies these have an effect small enough so that the value of ϵ is large. A similar effect can be obtained at higher frequencies by means of two resistance-capacitance coupled amplifiers connected as shown in Fig. 8, so that the output of the second is fed into the input of the first. Under these conditions the phase of the output voltage is such as to aid the grid of the input tube. The period of the oscillation is given roughly by $r_{g1}C_1 + r_{g2}C_2$. Using ordinary tubes these oscillators can be made to operate at frequencies up to 1,000,000 cycles.

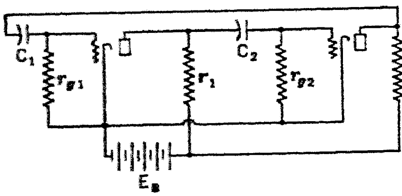


FIG. 8. The Usual Multivibrator Circuit

DISTORTIONLESS OSCILLATORS. As indicated under Non-linear Theory of Oscillators, p. 7-83, it is generally necessary for a stable oscillator to operate in such a way that an increase in grid-voltage amplitude would cause an increase in limiting-type distortion so that the voltage developed in the plate circuit produces too small a voltage in the grid circuit to maintain that voltage in the plate circuit. And similarly a decrease in grid-voltage amplitude would produce such a relatively increased voltage in the plate circuit that a larger voltage would be produced in the grid circuit, tending to return the grid-voltage amplitude to its original value. Under these conditions, the stable grid-voltage amplitude is such that the voltage developed in the plate circuit is just sufficient to maintain that stable grid voltage in the grid circuit. In other words, the voltage gain at the oscillating frequency must decrease as the amplitude increases and must increase as the amplitude decreases. The greater the magnitude of this dependence of gain upon amplitude, the greater the stability of the oscillator. In the ordinary oscillator, where this dependence of gain upon amplitude is accomplished through amplitude limiting, the greater the dependence, the greater the distortion. Consequently, it is generally true that an increase in stability of an oscillator is accompanied by an increase in distortion.

In any oscillator where the gain of the amplifier part of the circuit is controlled by the average amplitude, averaged over one or more cycles, rather than by the relatively instantaneous limiting action of the ordinary oscillator it is possible to operate the amplifier in that part of its characteristic where distortion is negligible. One such oscillator is the resistance-capacitance oscillator.

R-C OSCILLATOR. The resistance-capacitance tuned oscillator is shown in Fig. 9. The bridge circuit at the left of the circuit is a somewhat unbalanced Wien bridge at the oscillating frequency, $1/2\pi\sqrt{r_1r_2C_1C_2}$. The unbalance is such that, were r_b to be somewhat increased, the bridge would be balanced. Actually, r_a and r_b are so chosen that the unbalance voltage which is introduced in the grid circuit of the first amplifier tube is just sufficient when amplified to produce the required bridge suppy voltage between A and B to maintain that magnitude of unbalance voltage. If the gain of the amplifier were to increase from its normal value, increasing the voltage between A and B, the added heat developed in the lamp would cause r_b to increase, causing the bridge to become more nearly balanced and thus decreasing the input voltage to the amplifier to compensate for the increased gain. In the opposite fashion, a decrease in gain causes the bridge to be less nearly balanced. Thus such an oscillator is amplitude stable at the frequency which would balance the bridge. If $r_1C_1 = r_2C_2$, the maximum positive balance voltage appears at this frequency. Therefore, this oscillator operates stably at only one frequency. The gain control is averaged over several cycles owing to the thermal lag of the resistance change of the lamp. By means of ganged condensers for C_1 and C_2, the fre-

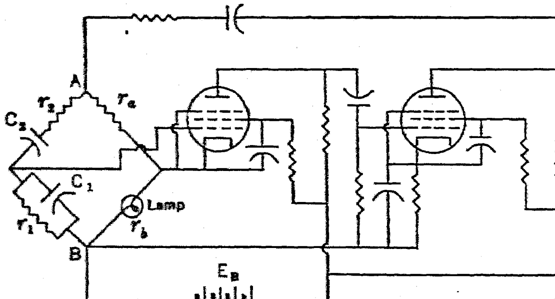


FIG. 9. Resistance-capacitance Tuned Oscillator

quency of such an oscillator can be conveniently varied. Oscillators of this type can be designed for frequencies well below the audible limit to several hundred kilocycles.

OSCILLATIONS OF GAS-FILLED TUBES. The gas-filled tube characteristic differs from the vacuum tube in one outstanding particular. The plate resistance varies from a very high value to a very low value almost discontinuously. Such a characteristic does not lend itself to the production of sinusoidal oscillations. However, it is highly advantageous for the production of almost discontinuous wave forms, the most important of which is the so-called sawtooth wave used for linear scanning in cathode-ray oscillographs.

The circuit of such an oscillator is shown in Fig. 10*a*. The condenser C is charged through the resistance r until the voltage across C is equal to the critical starting voltage of the plate circuit, the voltage at which the resistance suddenly decreases to a low value. The condenser is then discharged until the voltage across it equals the critical stopping voltage, the voltage at which the plate resistance returns to its high value again. This stopping voltage is always less than the starting voltage. The cycle then repeats itself as shown in Fig. 10*b*. These oscillators can be operated fairly satisfactorily up to 15,000 cycles, being limited by the ionization and deionization time of the gas. Above this value multivibrators must be used for linear sweep circuits.

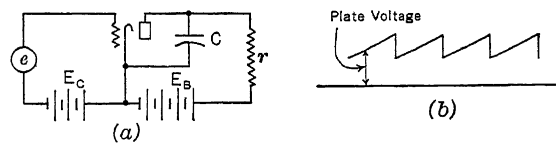


FIG 10. A Relaxation Oscillator, Using a Gaseous Tube, and Its Voltage Wave Form

These oscillators are easily synchronized with an external frequency by introducing a voltage at that frequency in the grid circuit as shown in the figure. As with vacuum-tube oscillators, these oscillators can be synchronized at a subharmonic of the synchronizing frequency.

BARKHAUSEN OSCILLATORS. In 1920 Barkhausen and Kurz discovered that, using certain tubes, oscillations a few centimeters in wavelength were produced when the grid supply voltage was highly positive and the plate was at or near the cathode potential Several theories of the operation of tubes under these conditions have been proposed but none agrees entirely with the experimental results. Practically all tubes which will oscillate under these conditions consist of cylindrical and coaxial cathodes, grids, and plates. Some investigators have succeeded in getting oscillations from other geometrical configurations, notably planes instead of cylinders.

Qualitatively, the explanation given is that electrons rapidly accelerated by means of the high grid potential pass through the grid as a result of their momentum and are turned back toward the grid, owing to its high potential, before they reach the plate. This causes each electron to set up a displacement current between the grid and the plate. A chaotic distribution of the phases of the oscillations of these electrons can be shown to be an unstable distribution. However, when these electrons oscillate in clouds the operation is stable so that the resultant current is not zero.

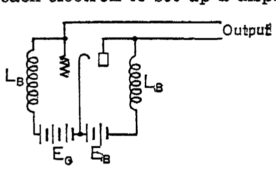


FIG. 11. The Barkhausen-Kurz Oscillator

The customary circuit for these oscillators is shown in Fig. 11. The power of the oscillations has been observed as high as 10 watts, the wavelengths go to 6 cm, and the efficiencies are from a fraction of 1 per cent to 7 per cent. The lowest efficiency and the lowest power usually occur at the shortest wavelengths.

OTHER SPECIAL OSCILLATORS. Other forms of oscillators use resonant cavities for the oscillating circuit element. These are treated in Section 4, article 8, on magnetrons, and article 22 of this section on cavity resonators.

21. ELECTROMECHANICAL OSCILLATORS

The frequency stability of mechanical vibrating systems is in general better than the frequency stability of electrical oscillatory circuits. For this reason mechanical vibrating systems are coupled to electrical circuits to give an electrical output of constant frequency.

TUNING-FORK OSCILLATORS. The circuit of a tuning-fork oscillator is shown in Fig. 12. This type of oscillator can be described as a refined buzzer. The resonant tuning fork corresponds to the oscillatory circuit, and the carbon button corresponds to the vacuum tube with the carbon as the plate circuit and the diaphragm as the grid. The feedback occurs through the transformer T. Although the electric circuit is tuned by means of the transformer inductances and the capacitance C, the harmonic content of the output is high. These oscillators are very convenient sources of audio-frequency voltage.

MAGNETOSTRICTION OSCILLATORS. When a body is placed in a magnetic field, stresses are produced within the body tending to distort it. Inversely, when a body is distorted, there is a change in the magnetic permeability. Magnetostriction is the name given to this effect. Many metals and alloys exhibit magnetostriction, but according to Ide it is most pronounced in alloys having 8–10 per cent chromium, 36–38 per cent nickel, and the remainder iron with about 1 per cent manganese to facilitate forging.

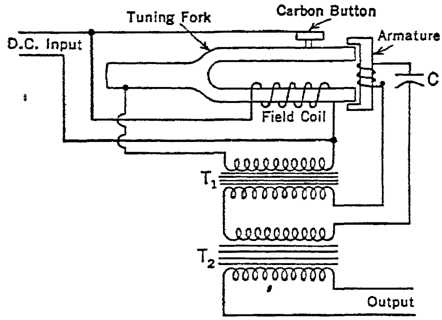


Fig. 12. The Tuning Fork Oscillator

When a rod of this material is magnetically polarized and placed in a coil carrying alternating current, it vibrates longitudinally at the frequency of the alternating current. If this frequency is the resonant frequency of the rod mechanically, the amplitude of the effect will be large even for very small currents in the coil. Thus it can be considered equivalent to a parallel tuned circuit coupled by means of the coil into the electric circuit. The resonant frequency is given by v/l, where v is the velocity of sound through the rod, about 4 km per sec, and l is the length. With the composition as above, and proper heat treatment and magnetization, the temperature coefficient is of the order of one part in a million per degree centigrade.

These devices are used as the oscillatory system of a vacuum-tube oscillator by means of a circuit such as that shown in Fig. 13. An additional coil carrying direct current may be necessary if the plate current is not sufficient to polarize the rod. Oscillations are more easily controlled when the relation between the direction of the coils is as shown, in contrast to that of the Hartley circuit (Fig. 1*a*). The positive feedback is obtained by means of the condenser C so that the electrical analogy is more nearly like the tuned-plate-tuned-grid oscillator shown in Fig. 1*d*. The rod is clamped in the middle (a node of its mechanical vibration), and the ends are free to vibrate.

The frequency stability of the magnetostriction oscillator compares favorably with that of the quartz crystal oscillators. The lower limit of frequency is determined by the practical limit of the length of the rod. The high-frequency limit is due to the magnetic skin effect of the rod. However, the harmonics of these high frequencies can be filtered out of the plate circuit so that frequencies of several million cycles can be obtained from these oscillators.

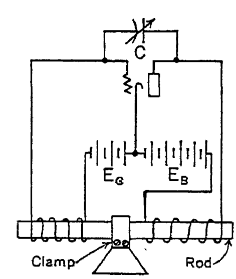


Fig. 13. The Magnetostriction Oscillator

PIEZOELECTRIC CRYSTAL OSCILLATORS. With respect to frequency, the most stable oscillators are oscillators controlled by piezoelectric crystals. Crystal oscillators are used in practically all transmitters as master oscillators. A crystal oscillator, used by the Bureau of Standards for the broadcast of standard frequency signals, gave a frequency stability of better than 1.5 parts in a million over a period of a year, and its stability was better than 2 parts in 100,000,000 over a period of several hours. By synchronizing oscillators to subharmonics of these standards, clocks can be driven which are considerably more accurate than pendulum clocks.

Crystal oscillators are applications of the piezoelectric effect, which is a means of coupling a mechanical motion to an electric circuit. When a strain is produced in a piezoelectric material, electric charges appear on its surface. Conversely, when an electric field is produced between surfaces of a piezoelectric material, a stress appears in the material. Thus if an alternating electromotive force is applied between two surfaces of the material, the material will vibrate, and if the material vibrates it will set up a displacement current between these surfaces.

This effect has been observed in many crystals, the most important of which are: quartz, tourmaline, and Rochelle salt (sodium potassium tartrate). Quartz and tourmaline have been used in oscillators, quartz crystals dominating the field. Tourmaline is more expensive than quartz but has the advantage that it can be ground to smaller sizes to produce oscillators having higher frequencies.

For use in conjunction with electric circuits, these crystals are cut in slabs, the geometry of which bears certain relations to the geometry of the crystal structure. These slabs are mounted in crystal holders, which consist essentially of two parallel conducting plates for the purpose of making an electrical connection with the surfaces of the crystal. An alternating emf set up between these plates causes a current to flow because of the piezoelectric properties of the crystal. When the frequency of this emf is equal to the mechanical resonant frequency of the crystal, the conductivity between the plates is maximum. This resonant frequency depends upon the dimensions of the slab and upon the relation between the geometry of the slab and the geometry of the crystal structure. The frequency of these resonators is therefore limited by the practical limitations of size of the slab. The frequency limits are from a few kilocycles to a few megacycles. (See Section 13, articles 32-34.)

For the analysis of electric circuits containing crystals it is convenient to replace the crystal by its equivalent electric circuit. This equivalent circuit is simply a series resonant circuit containing resistance, capacitance, and inductance paralleled by the capacitance of the holder. Because of the sharpness of tuning of this circuit, it is impossible actually to construct the electrical equivalent circuit.

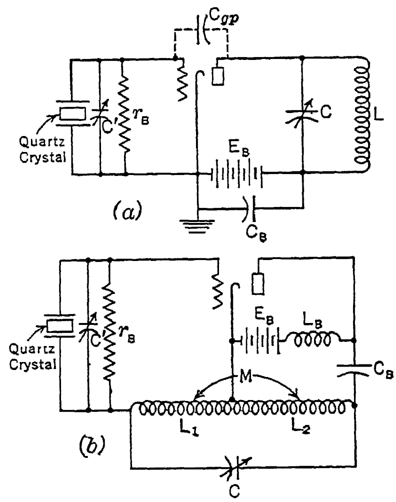


FIG. 14. Piezoelectric Crystal Oscillator Circuits

There are several oscillator circuits employing crystals, two of which are shown in Fig. 14. Circuit (*a*) makes use of the grid to plate interelectrode capacitance for feedback, as in the tuned-plate tuned-grid oscillator circuit (Fig. 1*d*); circuit (*b*) uses inductive feedback as in the Hartley oscillator (Fig. 1*a*). As in Fig. 1, the capacitances, C_B, are by-pass condensers, and the inductance, L_B, is a choke coil arranged to prevent the alternating current from passing through the d-c supply source, E_B. The tuned circuits containing L and C [in (*b*) $L = L_1 + L_2 + 2M$] are tuned to essentially the resonant frequency of the crystal.

The grid leaks r_B, as in ordinary vacuum-tube oscillators, furnish the operating bias for the tubes. The grid leak is in this case limited by an additional factor, namely, that the a-c current through the crystal, which is controlled by this resistance, must not exceed the safe operating value for the crystal. For low-frequency crystals this limit is 100 ma; for crystals in the megacycle range it reduces to 50 ma. Above this limit the crystal may vibrate violently enough to shatter itself. Since the d-c current through the grid leak is of the same order of magnitude as the a-c current through the crystal, an estimate of the value of this current can readily be obtained from the d-c grid current. In addition to this limitation the correct value for the grid leak is governed by the operating bias for the tube and varies from 10,000 ohms for high-μ tubes to 50,000 or more for low-μ tubes.

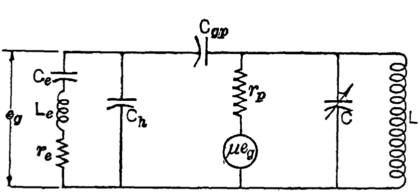


FIG. 15. The Equivalent Circuit Corresponding to Fig. 14*a*

The circuit shown in Fig. 14*a* is the most popular crystal-controlled oscillator circuit. Wheeler has analyzed the equivalent of this circuit (Fig. 15) by the method of van der Pol for the theory of non-linear oscillators. He obtained certain criteria for the frequency to be dependent primarily upon the resonant frequency of the crystal. First, the Q of the plate circuit tuning coil, i.e., the L to r ratio, should be as small as is consistent with stable oscillation. Second, the plate resistance of the tube should be as high as is consistent with stable oscillation. Third, C_h, which includes the capacitance C' and the grid to filament tube capacitance as well as the crystal holder capacitance, should approach but not exceed $\mu - 1$ times C_{gp}. Fourth, the plate tuning capacitance, C, should be adjusted to give maximum plate current.

The plate resistance can be made large by the choice of the tube and by decreasing the operating plate voltage. The capacitance C_h is usually made enough smaller than C_{gp}

$(\mu - 1)$ so that C' may be used as a fine adjustment of the frequency. The resonant frequency of the resonant circuit in the plate circuit must be above the resonant frequency of the crystal, while, if this resonant circuit has too high a frequency the plate load impedance is so low that oscillations cannot exist. As the capacitance, C, is varied from too low a value toward that value at which the resonant frequency of the plate circuit is the same as that of the crystal, the circuit at first fails to oscillate, then feeble oscillations start which increase steadily until just before the critical frequency the amplitude of oscillation rapidly drops to zero. The total variation in frequency of oscillation throughout this adjustment is 2 to 5 parts in 10,000. The optimum adjustment is for C to have a value just less than that for critical frequency.

Without some special means of temperature control, the resonant frequency of the crystal itself changes. The temperature coefficients of crystals vary, depending upon the relation between the geometry of the slab and the geometry of the crystal structure, from 1 part in 10,000 to 1 part in a million or less per degree centigrade. To minimize the variation in frequency due to the temperature, elaborate temperature-controlled ovens are used to maintain the temperature constant to better than 0.01 deg cent. For frequency control of an ordinary transmitter, ovens capable of maintaining temperatures to within 0.1 deg are sufficiently accurate.

A crystal oscillator may be designed to supply as much as 10 watts of output power at a plate efficiency of 30 to 60 per cent. The plate current and power output are limited by the current through the crystal. The usual d-c plate voltage is from 200 to 400 volts.

Because of the low power output and in order to prevent feedback into the crystal, a buffer amplifier must be used between these oscillators and the place where the power is to be applied, especially when the load is variable as in modulation.

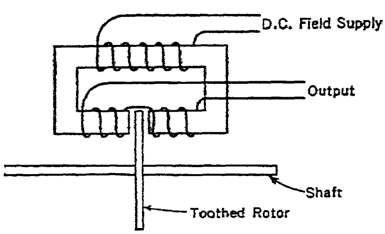


FIG. 16. Alexanderson Alternator

ALTERNATORS. Alternators are used primarily to produce a-c power at the low frequencies employed in power engineering. However, before the advent of high-power vacuum tubes, comparatively high frequencies for use in communication were obtained from specialized alternators. Two types of alternators were extensively employed for this purpose.

Alexanderson Alternators. The schematic diagram of an Alexanderson alternator is shown in Fig. 16. The toothed wheel is rotated between the poles of the magnet energized by a d-c field coil. This changes the magnetic flux density through the output coil periodically, setting up alternating currents in the load. The frequency of these alternators is limited to about 200 kc by the obvious mechanical limitations of construction and operation as well as by losses in the iron core and teeth.

Goldschmidt Alternators. The Goldschmidt alternators operated on a different principle. In a low-frequency single-phase alternator, the electromotive force developed in the armature contains not only the fundamental frequency but also the odd harmonics. Similarly, the field contains the even harmonics. Ordinarily, the reactances of the circuits are so arranged that the impedances to the fundamental and to the harmonics present are relatively high. The Goldschmidt alternators are designed so that the impedance at the frequencies of these harmonics is low. Then, by means of filters, one of the higher harmonics, usually the fourth harmonic, is selected for use. The frequency is limited here to about the same value by the same factors as in the Alexanderson alternator.

SPARK-GAP OSCILLATORS. Most of the early radio telegraph transmitters were damped-wave transmitters. These damped-wave oscillations were produced by means of spark-gap oscillators. The circuit of one of these oscillators is shown in Fig. 17. When the instantaneous voltage across the spark gap due to the audio-frequency generator exceeds the breakdown potential of the gap, a sudden rush of current shock-excites the oscillatory circuit consisting of the inductance L and the capacitance C, setting up an oscillatory current which is damped out by the resistance of the inductor, L, and by the load. The low resistance of the discharging arc is negligible until the voltage across it drops below some comparatively small value. This occurs twice during each cycle of the audio frequency. The most prominent frequency

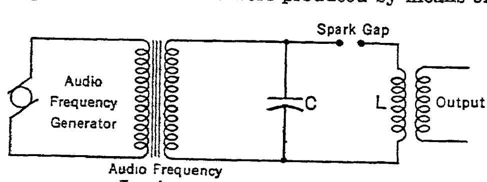


FIG. 17. Spark-gap Oscillator

in the output is the natural frequency of the resonant circuit, i.e., $\frac{1}{2\pi\sqrt{LC}}$. The output may be considered a carrier of this frequency modulated by an audio having a fundamental of twice the frequency of the generator voltage. There are so many harmonics of the audio that the bandwidth necessary for this type of transmission is very wide. For this reason, the ordinary use of spark transmitters is prohibited by law. However, many emergency stand-by transmitters on ships are of this type. The audio-frequency generator is frequently a buzzer using the transformer primary as the coil so that these transmitters may be operated by means of a battery.

Spark-gap transmitters are actively used at the present time for many industrial applications, chiefly to supply induction furnaces. The spark gaps are usually not air gaps, but a discharge in mercury vapor alone or with hydrogen. Such gaps can deliver high power at high efficiency, and they compare favorably with vacuum-tube oscillators for this purpose.

BIBLIOGRAPHY

Andrew, Victor J., The Adjustment of the Multivibrator for Frequency Division, *Proc. I.R.E.*, Vol. 19, 1911 (1934).

Fay, C. E., and A. L. Samuel, Vacuum Tubes for Generating Frequencies above 100 Megacycles, *Proc. I.R.E.*, Vol. 23, 199 (1935).

Hayasi, Tatuo, The Inner-grid Dynatron and the Duodynatron, *Proc. I.R.E.*, Vol. 22, 751 (1934).

Ide, John McDonald, Magnetostriction Alloys with Low Temperature Coefficients, *Proc. I.R.E.*, Vol. 22, 177 (1934).

Kelster, Frederick A., Generation and Utilization of Ultra-short Waves in Radio Communication, *Proc. I.R.E.*, Vol. 22, 1335 (1934).

Lapham, E. G., A 200-kilocycle Piezo Oscillator, *Bur. Standards J. Research*, Vol. 11, 59 (1933).

Llewellyn, F. B., Constant-frequency Oscillators, *Proc. I.R.E.*, Vol. 19, 2063 (1931).

Llewellyn, F. B., Note on Vacuum-tube Electronics at Ultra High Frequencies, *Proc. I.R.E.*, Vol. 23, 112 (1935).

MacKinnon, K. A., Crystal Control Applied to the Dynatron Oscillator, *Proc. I.R.E.*, Vol. 20, 1689 (1932).

McIlwain, Knox, and J. G. Brainerd, *High-frequency Alternating Currents*. John Wiley (1939).

Miles, L. D., A New Source of Kilocycle Kilowatts, *Elec. Engg.*, Vol. 305, 54 (1935).

Moore, W. H., Electron Oscillations without Tuned Circuits, *Proc. I.R.E.*, Vol. 22, 1021 (1934).

Morrison, W. A., A High Precision Standard of Frequency, *Proc. I.R.E.*, Vol. 17, 1103 (1929).

Peterson, E., J. G. Kreer, and L. A. Ware, Regeneration Theory and Experiment, *Proc. I.R.E.*, Vol. 22, 1191 (1934).

Prince, D. C., and F. B. Vogdes, *Vacuum Tubes as Oscillation Generators*. General Electric Review, Schenectady, N. Y. (1929).

Schneider, E. G., Radar, *Proc. I.R.E.*, Vol. 34, 528 (1946).

Terman, F. E., *Radio Engineering*. McGraw-Hill (1932).

Terman, F. E., R. R. Buss, W. R. Hewlett, and F. C. Cahill, Some Applications of Negative Feedback with Particular Reference to Laboratory Equipment, *Proc. I.R.E.*, Vol. 10,649 (1939).

Thompson, B. J., and G. M. Rose, Jr., Vacuum Tubes of Extremely Small Dimensions for Use at Extremely High Frequencies, *Proc. I.R.E.*, Vol. 21, 1707 (1933).

Thompson, B. J., and P. D. Zottu, An Electron Oscillator with Plane Electrodes, *Proc. I.R.E.*, Vol. 22, 1374 (1934).

van der Pol, Balth, The Non-linear Theory of Electric Oscillations, *Proc. I.R.E.*, Vol. 22, 1051 (1934).

Wheeler, L. P., Analysis of a Piezoelectric Oscillator Circuit, *Proc. I.R.E.*, Vol. 19, 627 (1931).

22. CAVITY RESONATORS

By I. G. Wilson and J. P. Kinzer

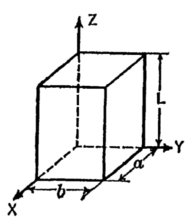
A cavity resonator is a section of dielectric completely surrounded by a metallic shell. In many ways, its performance is analogous to a resonant R, L, C low-frequency circuit. However L, C, and R can no longer serve as basic in the consideration of cavity resonators, because of the inability to define inductance and capacitance uniquely. (See reference 1.) In fact it is possible to find only two such quantities which describe the properties of a cavity resonator.

The first of these is the resonant frequencies (or wavelengths), defined as those values of f (or λ) which result in the boundary conditions being satisfied. With each resonance there is associated a particular standing-wave pattern of the electromagnetic fields. These have been called "eigenvalues" or "eigentones," but the term "normal modes" is now in general use.

The second is the quality factor, Q, defined by

$$Q = 2\pi \frac{\text{Energy stored}}{\text{Energy lost per cycle}} \tag{5}$$

Table 1. Formulas for Cavity Resonators

Type of Cavity and Coordinate System	Mode	Field Equations *	Definitions	Restrictions on l, m, n
Rectangular Prism 	*TM*	$E_x = \sqrt{\frac{\mu}{\epsilon}}\frac{k_1k_3}{k^2}\cos k_1x \sin k_2y \sin k_3z$ $E_y = \sqrt{\frac{\mu}{\epsilon}}\frac{k_2k_3}{k^2}\sin k_1x \cos k_2y \sin k_3z$ $E_z = -\sqrt{\frac{\mu}{\epsilon}}\frac{k_1^2+k_2^2}{k^2}\sin k_1x \sin k_2y \cos k_3z$ $H_x = -\frac{k_2}{k}\sin k_1x \cos k_2y \cos k_3z$ $H_y = \frac{k_1}{k}\cos k_1x \sin k_2y \cos k_3z$ $H_z = 0$	$k_1 = \frac{l\pi}{a} \quad k_2 = \frac{m\pi}{b} \quad k_3 = \frac{n\pi}{L}$ $k^2 = k_1^2 + k_2^2 + k_3^2 \quad \lambda = \frac{2\pi}{k}$	$l > 0$ $m > 0$
	TE	$E_x = -\sqrt{\frac{\mu}{\epsilon}}\frac{k_2}{k}\cos k_1x \sin k_2y \sin k_3z$ $E_y = \sqrt{\frac{\mu}{\epsilon}}\frac{k_1}{k}\sin k_1x \cos k_2y \sin k_3z$ $E_z = 0$ $H_x = \frac{k_1k_3}{k^2}\sin k_1x \cos k_2y \cos k_3z$ $H_y = \frac{k_2k_3}{k^2}\cos k_1x \sin k_2y \cos k_3z$ $H_z = -\frac{k_1^2+k_2^2}{k^2}\cos k_1x \cos k_2y \sin k_3z$	l, m, n = integral indices identifying the modes. May assume the value zero, subject to restrictions given in adjoining column	$l + m > 0$ $n > 0$

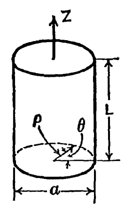
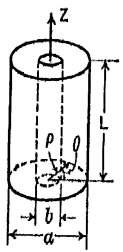
Circular Cylinder 	TM	$E_\rho = -\sqrt{\frac{\mu}{\epsilon}}\frac{k_3}{k}J'_l(k_1\rho)\cos l\theta \sin k_3z$ $E_\theta = \sqrt{\frac{\mu}{\epsilon}}\,l\,\frac{k_3}{k}\frac{J_l(k_1\rho)}{k_1\rho}\sin l\theta \sin k_3z$ $E_z = \sqrt{\frac{\mu}{\epsilon}}\frac{k_1}{k}J_l(k_1\rho)\cos l\theta \cos k_3z$ $H_\rho = -l\frac{J_l(k_1\rho)}{(k_1\rho)}\sin l\theta \cos k_3z$ $H_\theta = -J'_l(k_1\rho)\cos l\theta \cos k_3z$ $H_z = 0$	$k_1 = \frac{2r_{lm}}{a} \quad k_3 = \frac{n\pi}{L}$ $k^2 = k_1^2 + k_3^2 \quad \lambda = \frac{2\pi}{k}$	$m > 0$
	TE	$E_\rho = -\sqrt{\frac{\mu}{\epsilon}}\,l\,\frac{J_l(k_1\rho)}{k_1\rho}\sin l\theta \sin k_3z$ $E_\theta = -\sqrt{\frac{\mu}{\epsilon}}J'_l(k_1\rho)\cos l\theta \sin k_3z$ $E_z = 0$ $H_\rho = \frac{k_3}{k}J'_l(k_1\rho)\cos l\theta \cos k_3z$ $H_\theta = -l\frac{k_3}{k}\frac{J_l(k_1\rho)}{k_1\rho}\sin l\theta \cos k_3z$ $H_z = \frac{k_1}{k}J_l(k_1\rho)\cos l\theta \sin k_3z$	l, m, n = defined as for rectangular prism r_{lm} = m^{th} zero of $J_l(x)$ for TM modes r_{lm} = m^{th} zero of $J'_l(x)$ for TE modes	$m > 0$ $n > 0$

Table 1—*Continued*

Type of Cavity and Coordinate System	Mode	Field Equations *	Definitions	Restrictions on l, m, n
Full Coaxial  $\frac{b}{a} = \eta$	TM	Same as for circular cylinder, but substitute: $Z_l(k_1\rho)$ for $J_l(k_1\rho)$; $Z'_l(k_1\rho)$ for $J'_l(k_1\rho)$ where $Z_l(k_1\rho) = J_l(k_1\rho) - AY_l(k_1\rho)$; $Z'_l(k_1\rho) = J'_l(k_1\rho) - AY'_l(k_1\rho)$	Same as circular cylinder, except: $r_{lm} = m^{th}$ zero of $[J_l(\eta x)Y_l(x) - J_l(x)Y_l(\eta x)]$; $A = \frac{J_l(r_{lm})}{Y_l(r_{lm})}$	$m > 0$. Special case of TM 0, 0, n mode, with $r_{lm} = 0$
	TE		$r_{lm} = m^{th}$ zero of $[J'_l(\eta x)Y'_l(x) - J'_l(x)Y'_l(\eta x)]$; $A = \frac{J'_l(r_{lm})}{Y'_l(r_{lm})}$	$m > 0$; $n > 0$

Sources: Hansen, *Jnl. App. Phys.*, Vol. 9, p. 654; Borgnis, *Hochf: tech u. Elek: Akus.*, Vol. 56, p. 47; Borgnis, *Ann. d. Phys.*, Vol. 35, p. 359; Barrow & Mieher, *Proc. I.R.E.*, Vol. 28, p. 184.

* The time factor has been omitted; the E-field is in time quadrature with the H-field, with $\omega = ck$.

Table 1—*Continued*

Cavity	Mode	Normal Wavelengths	Approximation for Total Number of Modes (*TE* and *TM*) Having $\lambda > \lambda_0$	Formulas for $Q \frac{\delta}{\lambda}$	Definitions
Rectangular Prism	*TM*	$\lambda = \dfrac{2}{\sqrt{\left(\frac{l}{a}\right)^2 + \left(\frac{m}{b}\right)^2 + \left(\frac{n}{L}\right)^2}}$	$N = 8.38 \dfrac{V}{\lambda_0^3} - \dfrac{P}{\lambda_0}$ $V = abL$ $P = a + b + L$	$\dfrac{abL}{4} \cdot \dfrac{(p^2 + q^2)(p^2 + q^2 + r^2)^{1/2}}{p^2 b(a + L) + q^2 a(b + L)}$ $n > 0$ $\dfrac{abL}{2} \cdot \dfrac{(p^2 + q^2)^{3/2}}{p^2 b(a + 2L) + q^2 a(b + 2L)}$ $n = 0$	$p = \dfrac{l}{a}$ $q = \dfrac{m}{b}$ $r = \dfrac{n}{L}$
	TE	Same as *TM* modes		$\dfrac{abL}{4} \cdot \dfrac{(p^2 + q^2)(p^2 + q^2 + r^2)^{3/2}}{aL[p^2r^2 + (p^2 + q^2)^2] + bL[q^2r^2 + (p^2 + q^2)^2] + abr^2(p^2 + q^2)}$ $lm > 0$ $\dfrac{abL}{2} \cdot \dfrac{(q^2 + r^2)^{3/2}}{q^2 L(b + 2a) + r^2 b(L + 2a)}$ $l = 0$ $\dfrac{abL}{2} \cdot \dfrac{(p^2 + r^2)^{3/2}}{p^2 L(a + 2b) + r^2 a(L + 2b)}$ $m = 0$	
Circular Cylinder	*TM*	$\lambda = \dfrac{2}{\sqrt{\left(\frac{2r_{lm}}{\pi a}\right)^2 + \left(\frac{n}{L}\right)^2}}$ $(fa)^2 = \left(\dfrac{cr_{lm}}{\pi}\right)^2 + \left(\dfrac{cn}{2}\right)^2 \left(\dfrac{a}{L}\right)^2$ $c = \sqrt{\mu\epsilon}$ = velocity of electromagnetic waves in dielectric f = frequency	$N = 4.38 \dfrac{V}{\lambda_0^3} +$ $0.09 \dfrac{S}{\lambda_0^2}$ $V = \dfrac{\pi a^2 L}{4}$ $S = \pi a L$	$\dfrac{r_{lm}}{2\pi} [1 + p^2R^2]^{1/2} \cdot \dfrac{1}{1 + R}$ $n > 0$ $\dfrac{r_{lm}}{\pi} \dfrac{1}{2 + R}$ $n = 0$	$R = \dfrac{a}{L}$ $p = \dfrac{n\pi}{2r_{lm}}$
	TE			$\dfrac{r_{lm}}{2\pi} [1 + p^2R^2]^{3/2} \dfrac{1 - \left(\dfrac{l}{r_{lm}}\right)^2}{1 + p^2R^3 + p^2(1 - R)R^2 \left(\dfrac{l}{r_{lm}}\right)^2}$	

Table 1—*Continued*

Cavity	Mode	Normal Wavelengths	Approximation for Total Number of Modes (*TE* and *TM*) Having $\lambda > \lambda_0$	Formulas for $Q\frac{\delta}{\lambda}$	Definitions
Full Co-axial	*TM*	Same form as for cylinder r_{lm} has different values	$N \approx 4.4 \frac{V}{\lambda_0^3}$ With some doubt as to value of the coefficient	$\frac{r_{lm}}{2\pi}[1 + p^2R^2]^{1/2} \frac{(1 - \eta^2H')}{(1 + \eta H') + R(1 - \eta^2H')}$ $n > 0$ $\frac{r_{lm}}{\pi} \cdot \frac{(1 - \eta^2H')}{2(1 + \eta H') + R(1 - \eta^2H')}$ $n = 0$ These expressions are not valid for small η when $l = 0$	$R = \frac{a}{L}$ $p = \frac{n\pi}{2r_{lm}}$ $H' = \left[\frac{Z'_l(\eta r_{lm})}{Z'_l(r_{lm})}\right]^2$ $H = \left[\frac{Z_l(\eta r_{lm})}{Z_l(r_{lm})}\right]^2$
	TE			$\frac{r_{lm}}{2\pi} \cdot \frac{[1 + p^2R^2]^{3/2}M}{(1 + \eta H) + p^2R^2 \frac{l^2}{r_{lm}^2}\left(1 + \frac{H}{\eta}\right) + p^2R^3M}$ where $M = \left(1 - \frac{l^2}{r_{lm}^2}\right) - \eta^2H\left(1 - \frac{l^2}{\eta^2 r_{lm}^2}\right)$	

In calculating λ, the assumption is made that the walls of the cavity have perfect conductivity. When Q is calculated, the assumption is made that λ is unchanged; that is, the fields that actually exist are those calculated on the basis of perfectly conducting walls. Since the Q's are generally high, the approximation is extremely good. (See reference 2.)

Since the energy is stored in the cavity volume, while the energy is lost in the walls, to obtain a high Q, the resonator should have a large ratio of volume to surface area. For this reason, cylinders, prisms, or spheres will in general have better Q's than cavities with re-entrant portions, or coaxial structures.

In computing the resonant frequencies and Q values of cavity resonators, solutions of the field equations involve, for the rectangular prism, circular functions; for the perfect cylinder, Bessel functions of the first kind; and, for two coaxial cylinders, Bessel functions of the first and second kinds. Only approximate solutions have been derived for cavities involving re-entrant sections.

MODES. By fundamental and general consideration, the modes in every cavity resonator, regardless of its shape, are infinite in number and more closely spaced as the frequency increases. The total number, N, of these having a resonant frequency less than f is given approximately by:

$$N = \frac{8\pi}{3c^3} Vf^3 \tag{6}$$

where V = volume of cavity in cubic meters.
c = velocity of electromagnetic waves in the dielectric in meters per second.
f = frequency in cycles per second.

PRINCIPLE OF SIMILITUDE. Another theorem generally applicable to all cavities is the principle of similitude, stated as follows (see reference 3): A reduction in all the linear dimensions of a cavity resonator by a factor $1/m$ (if accompanied by an increase in the conductivity of the walls by a factor m) will reduce the wavelengths of the modes by the factor $1/m$. For high-Q cavities, the condition given in parentheses need not be considered.

MODES IN RIGHT CYLINDERS. In right cylinders (ends perpendicular to axis) the modes fall naturally into two groups, the transverse electric (TE) and the transverse magnetic (TM). In the TE modes, the electric lines everywhere lie in planes perpendicular to the cylinder axis, and in the TM modes the magnetic lines so lie. Further identification of a specific mode is accomplished by the use of indices.

THE *MS* FACTOR. With a cylinder restricted to a loss-free dielectric and a non-magnetic surface, the value of Q (quality factor) for each mode depends on the conductivity of the metallic surface, the frequency, and the ratio of a cross-sectional dimension to the length. The quantity $Q\delta/\lambda$, however, depends only upon the mode and shape and is referred to as the mode-shape (MS) factor.

In this expression, δ refers to skin depth (see reference 4) in meters $= (1/2\pi)(\sqrt{10^7\rho/f})$, λ is wavelength in meters in the dielectric $= c/f$, and ρ the resistivity in ohm-meters. The skin depth is a factor which recognizes the dissipation of energy in the walls and ends of the cylinder. With increase of resistivity of these surfaces the currents penetrate deeper and the resulting Q is lower.

FUNDAMENTAL FORMULAS. Expressions for standing-wave patterns and $Q\delta/\lambda$ are given in Table 1, for right rectangular, circular, and full coaxial cylinders. The mode indices are l, m, n following the notation of Barrow and Mieher (reference 5). In the rectangular prism they denote the number of half-wavelengths along the coordinate axes. For the other two cylinders they have an analogous physical significance with l related to the angular coordinate, m to the radial, and n to the axial.

In an elliptical cylinder, a further index is needed to distinguish between modes which differ only in their orientation with respect to the major and minor axes; these paired modes are termed even and odd and have slightly different resonant frequencies (see reference 6). In the circular cylinder they have the same frequency, a condition which is referred to as a degeneracy (in this case, double); that is, in the circular cylinder, odd and even modes are distinguishable only by a difference in their orientation within the cylinder with reference to the origin of the angular coordinate. In Table 1, the field expressions are given for the even modes; those for the odd modes are obtained by changing $\cos l\theta$ to $\sin l\theta$ and $\sin l\theta$ to $\cos l\theta$ everywhere.

The value of N in the table for the circular cylinder is based on counting this degeneracy as a single mode; counting even and odd modes as distinct will nearly double the value of N, thus bringing it into agreement with the general eq. (6).

In Table 1, the mks system of units is implied. The notation is in general accordance with that used in prior developments of the subject. For engineering applications, it is advantageous to reduce the results to units in ordinary use and to change the notation wherever this leads to a more obvious association of ideas. For these reasons, in what follows attention is confined to the right circular cylinder, with changes in units and notation as specified later.

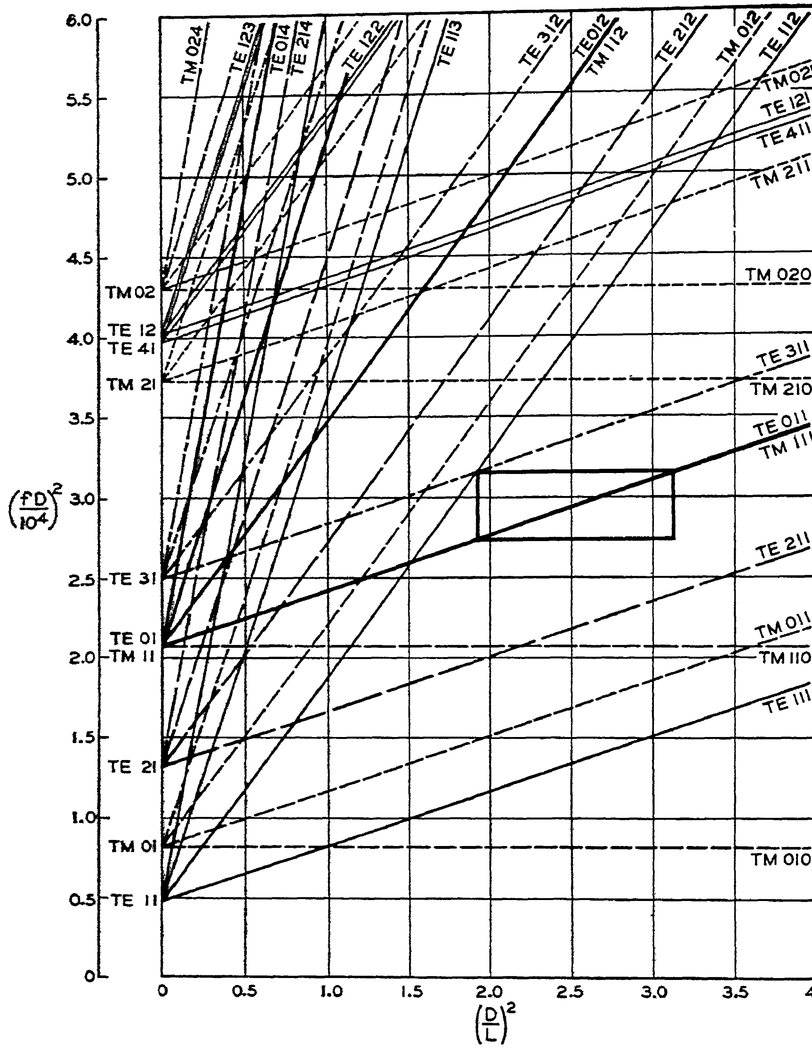


FIG. 18. Mode Chart for Circular Cylinder Cavity Resonator. D and L in inches; f in megacycles per second. The rectangle indicates the best operating region for the TE 011 mode.

THE MODE CHART. The formula relating the resonant frequency to the mode, shape, and dimensions may be written simply:

$$(fD)^2 = A + Bn^2\left(\frac{D}{L}\right)^2 \tag{7}$$

where f = frequency in megacycles per second.

D = diameter of cavity in inches.

L = length of cavity in inches.

A = a constant depending upon the mode. Values of A are given in Table 2 for the lowest 30 modes. Values of Bessel function roots are given in Table 3 for first 180 modes.

B = a constant depending upon the velocity of electromagnetic waves in the dielectric. For air at 25 deg cent and 60 per cent relative humidity, $B = 0.34799 \times 10^8$.

n = third index defining the mode, i.e., the number of half wavelengths along the cylinder axis.

Formula (7) represents a family of straight lines, when $(D/L)^2$ and $(fD)^2$ are used as coordinates, and leads directly to the easily constructed and highly useful mode chart of Fig. 18.

It will be noted from Table 2 that the *TE* 0*mn* and the *TM* 1*mn* modes have the same frequency of resonance. This is a highly important case of degeneracy. In the design of practical cavities it is necessary to take measures to eliminate this degeneracy, as the *TM* mode (usually referred to as the companion of its associated *TE* mode) introduces undesirable effects.

Table 2. Constants for Use in Computing the Resonant Frequencies of Circular Cylinders

$$(fD)^2 = \left(\frac{cr}{\pi}\right)^2 + \left(\frac{cn}{2}\right)^2 \left(\frac{D}{L}\right)^2$$

$$= A + Bn^2 \left(\frac{D}{L}\right)^2$$

$B = 0.34799 \times 10^8$

$c = 1.17981 \times 10^{10}$ in. per sec

Mode	r	A
TM 01	2.40483	0.81563×10^8
02	5.52008	4.2975
03	8.65373	10.5617
11	3.83171	2.0707
12	7.01559	6.9415
13	10.17347	14.5970
21	5.13562	3.7197
22	8.41724	9.9923
31	6.38016	5.7410
32	9.76102	13.4374
41	7.58834	8.1212
51	8.77148	10.8511
61	9.93611	13.9238
TE 01	3.83171	2.0707
02	7.01559	6.9415
03	10.17347	14.5970
11	1.84118	0.47810
12	5.33144	4.0088
13	8.53632	10.2770
21	3.05424	1.3156
22	6.70613	6.3426
23	9.96947	14.0175
31	4.20119	2.4893
32	8.01524	9.0606
41	5.31755	3.9879
42	9.28240	12.1520
51	6.41562	5.8050
61	7.50127	7.9359
71	8.57784	10.3772
81	9.64742	13.1265

Value of c is for air at 25 deg cent and 60 per cent relative humidity. D and L in inches; f in megacycles.

DESIGN OF HIGH-Q CAVITY IN *TE* 01n MODE. In many applications, a resonator is used in its fundamental (gravest) mode. However, when high values of Q are desired, it may be necessary to use a high-order mode. In this case, it is desirable to keep the volume the minimum because other modes can cause undesired responses and other deleterious effects. As shown by eq. (6), the total number of resonances is a function of the volume. Analysis of the problem leads to the conclusion that operation in the *TE* 01n mode (unimportant exceptions occur for values of $Q\delta/\lambda$ less than 1.2) gives the smallest volume for an assigned Q and also leads to specific values of n and D/L which give this result. In fact, for maximum Q per volume in the *TE* 01n mode,

$$(fD)^2 \frac{D}{L} = 3.11 \times 10^8 \tag{8}$$

which permits easy plotting on a mode chart of the locus of the operating points for best Q per volume ratio.

The mode-shape (*MS*) factor for the *TE* 01n modes may be expressed as follows:

$$\frac{Q\sqrt{f}}{10^6} = 2.77 \frac{\left[1 + 0.168 \left(\frac{D}{L}\right)^2 n^2\right]^{3/2}}{1 + 0.168 \left(\frac{D}{L}\right)^3 n^2} \tag{9}$$

This has been derived from Table 1 by combining terms which are a function of frequency and by assuming the conductivity of copper ($\rho = 1.7241 \times 10^{-8}$ – the International Standard value for copper) for the cylinder walls.

The relative Q's for several metals are: silver 1.03, copper 1.00, gold 0.84, aluminum 0.78, and brass 0.48. Therefore, a brass cavity will have about one-half of the Q of a similar copper cavity. Silver-plating a copper cavity will increase Q about 3 per cent. Experience shows that only 80 to 90 per cent of the theoretical Q can be realized. This should be taken account of in the design.

With the frequency and desired Q known, the dimensions of the cavity can be determined.

CAVITY COUPLINGS. To be useful the cavity must be coupled to external circuits. The coupling to all modes can be analyzed, at least qualitatively, from the field expressions of Table 1. The problem is to get the correct coupling to the desired mode and as little coupling as possible to all others. This may be obtained either by a loop or a probe at the end of a coaxial line or by an orifice connecting the cavity with a wave guide. For optimum coupling the plane of a loop must be perpendicular to the H lines; the axis of a probe must be collinear with the E lines; and the H lines in a wave guide feeding through an orifice must be parallel to the H lines in the cavity.

Since the electric field is zero everywhere at the boundary surface of the cavity for the TE 01n mode, coupling to it must be magnetic; a probe cannot be used. The location for

Table 3. Values of the Bessel Function Zero (r_{lm}) for the First 180 Modes in a Circular Cylinder Resonator

	r_{lm}	Mode *		r_{lm}	Mode		r_{lm}	Mode		r_{lm}	Mode
1	1.8412	*E* 1-1	46	13.0152	*M* 3-3	91	18.4335	*M* 10-2	136	22.6716	*E* 2-7
2	2.4048	*M* 0-1	47	13.1704	*E* 2-4	92	18.6374	*E* 6-4	137	22.7601	*M* 1-7
3	3.0542	*E* 2-1	48	13.3237	*M* 1-4	93	18.7451	*E* 12-2	138	22.7601	*E* 0-7
4	3.8317	*M* 1-1	49	13.3237	*E* 0-4	94	18.9000	*M* 14-1	139	22.9452	*M* 8-4
5	3.8317	*E* 0-1	50	13.3543	*M* 9-1	95	18.9801	*M* 5-4	140	23.1158	*M* 14-2
6	4.2012	*E* 3-1	51	13.5893	*M* 6-2	96	19.0046	*E* 9-3	141	23.2548	*E* 21-1
7	5.1356	*M* 2-1	52	13.8788	*E* 12-1	97	19.1045	*E* 17-1	142	23.2568	*M* 18-1
8	5.3176	*E* 4-1	53	13.9872	*E* 5-3	98	19.1960	*E* 4-5	143	23.2643	*E* 16-2
9	5.3314	*E* 1-2	54	14.115$\underline{5}$	*E* 8-2	99	19.4094	*M* 3-5	144	23.2681	*E* 7-5
10	5.5201	*M* 0-2	55	14.372$\underline{5}$	*M* 4-3	100	19.5129	*E* 2-6	145	23.2759	*M* 11-3
11	6.3802	*M* 3-1	56	14.475$\underline{5}$	*M* 10-1	101	19.554$\underline{5}$	*M* 8-3	146	23.5861	*M* 6-5
12	6.4156	*E* 5-1	57	14.5858	*E* 3-4	102	19.6159	*M* 1-6	147	23.7607	*E* 10-4
13	6.7061	*E* 2-2	58	14.7960	*M* 2-4	103	19.6159	*E* 0-6	148	23.8036	*E* 5-6
14	7.0156	*M* 1-2	59	14.8213	*M* 7-2	104	19.6160	*M* 11-2	149	23.8194	*E* 13-3
15	7.0156	*E* 0-2	60	14.8636	*E* 1-5	105	19.8832	*E* 13-2	150	24.0190	*M* 4-6
16	7.5013	*E* 6-1	61	14.9284	*E* 13-1	106	19.9419	*E* 7-4	151	24.1449	*E* 3-7
17	7.5883	*M* 4-1	62	14.9309	*M* 0-5	107	19.9944	*M* 15-1	152	24.2339	*M* 9-4
18	8.0152	*E* 3-2	63	15.2682	*E* 6-3	108	20.1441	*E* 18-1	153	24.2692	*M* 15-2
19	8.4172	*M* 2-2	64	15.2867	*E* 9-2	109	20.2230	*E* 10-3	154	24.2701	*M* 2-7
20	8.5363	*E* 1-3	65	15.5898	*M* 11-1	110	20.3208	*M* 6-4	155	24.2894	*E* 22-1
21	8.5778	*E* 7-1	66	15.7002	*M* 5-3	111	20.575$\underline{5}$	*E* 5-5	156	24.3113	*E* 1-8
22	8.6537	*M* 0-3	67	15.9641	*E* 4-4	112	20.7899	*M* 12-2	157	24.3382	*M* 19-1
23	8.7715	*M* 5-1	68	15.9754	*E* 14-1	113	20.8070	*M* 9-3	158	24.3525	*M* 0-8
24	9.2824	*E* 4-2	69	16.0378	*M* 8-2	114	20.8269	*M* 4-5	159	24.3819	*E* 17-2
25	9.6474	*E* 8-1	70	16.2235	*M* 3-4	115	20.9725	*E* 3-6	160	24.4949	*M* 12-3
26	9.7610	*M* 3-2	71	16.347$\underline{5}$	*E* 2-5	116	21.0154	*E* 14-2	161	24.5872	*E* 8-5
27	9.9361	*M* 6-1	72	16.4479	*E* 10-2	117	21.0851	*M* 16-1	162	24.9349	*M* 7-5
28	9.9695	*E* 2-3	73	16.4706	*M* 1-5	118	21.1170	*M* 2-6	163	25.0020	*E* 14-3
29	10.1735	*M* 1-3	74	16.4706	*E* 0-5	119	21.1644	*E* 1-7	164	25.008$\underline{5}$	*E* 11-4
30	10.1735	*E* 0-3	75	16.5294	*E* 7-3	120	21.1823	*E* 19-1	165	25.1839	*E* 6-6
31	10.5199	*E* 5-2	76	16.6982	*M* 12-1	121	21.2116	*M* 0-7	166	25.3229	*E* 23-1
32	10.7114	*E* 9-1	77	17.0038	*M* 6-3	122	21.2291	*E* 8-4	167	25.4170	*M* 16-2
33	11.0647	*M* 4-2	78	17.0203	*E* 15-1	123	21.4309	*E* 11-3	168	25.4171	*M* 20-1
34	11.0864	*M* 7-1	79	17.2412	*M* 9-2	124	21.641$\underline{5}$	*M* 7-4	169	25.4303	*M* 5-6
35	11.3459	*E* 3-3	80	17.3128	*E* 5-4	125	21.9317	*E* 6-5	170	25.4956	*E* 18-2
36	11.6198	*M* 2-3	81	17.6003	*E* 11-2	126	21.9562	*M* 13-2	171	25.5094	*M* 10-4
37	11.7060	*E* 1-4	82	17.6160	*M* 4-4	127	22.0470	*M* 10-3	172	25.5898	*E* 4-7
38	11.7349	*E* 6-2	83	17.7740	*E* 8-3	128	22.1422	*E* 15-2	173	25.7051	*M* 13-3
39	11.7709	*E* 10-1	84	17.7887	*E* 3-5	129	22.1725	*M* 17-1	174	25.7482	*M* 3-7
40	11.791$\underline{5}$	*M* 0-4	85	17.8014	*M* 13-1	130	22.2178	*M* 5-5	175	25.8260	*E* 2-8
41	12.2251	*M* 8-1	86	17.9598	*M* 2-5	131	22.2191	*E* 20-1	176	25.8912	*E* 9-5
42	12.3386	*M* 5-2	87	18.015$\underline{5}$	*E* 1-6	132	22.4010	*E* 4-6	177	25.9037	*M* 1-8
43	12.6819	*E* 4-3	88	18.0633	*E* 16-1	133	22.5014	*E* 9-4	178	25.9037	*E* 0-8
44	12.8265	*E* 11-1	89	18.0711	*M* 0-6	134	22.5827	*M* 3-6	179	26.1778	*E* 15-3
45	12.9324	*E* 7-2	90	18.2876	*M* 7-3	135	22.6293	*E* 12-3	180	26.2460	*E* 12-4

* Nomenclature after Barrow and Mieher, Natural Oscillations of Electrical Cavity Resonators, *Proc. I.R.E.*, April 1940, p. 184.

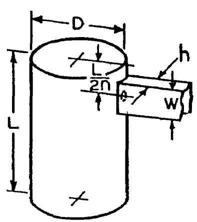
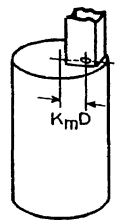
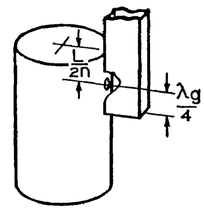
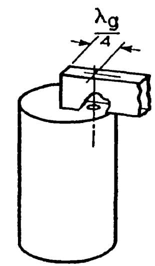
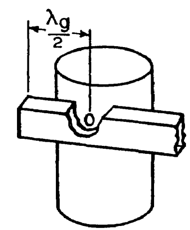
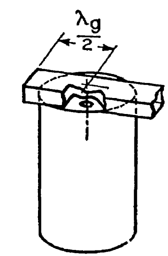
M modes take zeros of $J_l(x)$; E modes take zeros of $J_l'(x)$. Number directly following E or M is l; number after hyphen is number of root.

Values less than 16.0 are abridged from six-place values and are believed to be correct; values more than 16.0 are abridged from five-place values and may be in error by one unit in fourth decimal place. Underlined $\underline{5}$ in fourth place indicates that higher value is to be used in rounding off to fewer decimals.

maximum coupling is on the side of the cavity, an odd number of quarter-guide wavelengths from the end, or on the end about halfway (48 per cent) out from the center to the edge. Correct orientation requires the axis of a loop to be parallel to the axis of the cylinder for side-wall feed and to be perpendicular to the cylinder axis for end feed. Wave-guide orientation is shown in Table 4.

The theory of coupling loops and orifices is not at present precise enough to yield more than approximate dimensions. On the basis of rather severely limiting assumptions, coupling formulas for a round hole connecting a rectangular wave guide and a TE 01n

Table 4. Orifice Coupling of Wave Guide (in *TE* 10 Mode) to Cylindrical Cavity (in *TE* 01n Mode)

	CASE 1A	CASE 1B	CASE 2A	CASE 2B	CASE 3A	CASE 3B
COUPLING METHOD						
CIRCULAR ORIFICE	$\frac{\Delta f}{f} = -K_c \frac{\lambda^2 d^3}{D^4 L}$ $W_a = K_w \frac{\lambda^2 d^6}{\lambda_g w h D^4 L}$	$\frac{\Delta f}{f} = -K_c \frac{n^2 \lambda^2 d^3}{D^2 L^3}$ $W_a = K_w \frac{n^2 \lambda^2 d^6}{\lambda_g w h D^2 L^3}$	$\frac{\Delta f}{f}$ SAME AS 1A $W_a = K_w \frac{\lambda_g \lambda^2 d^6}{w^3 h D^4 L}$	$\frac{\Delta f}{f}$ SAME AS 1B $W_a = K_w \frac{n^2 \lambda_g \lambda^2 d^6}{w^3 h D^2 L^3}$	$\frac{\Delta f}{f}$ SAME AS 1A W_a SAME AS 1A	$\frac{\Delta f}{f}$ SAME AS 1B W_a SAME AS 1B

CONSTANTS

	1A K_c	1A K_w	1B K_c	1B K_w	1B K_m	2A K_w	2B K_w
TE 01n	0.316	1.322	0.1107	0.464	0.2403	0.331	0.1159
TE 02n	1.058	4.43	0.1995	0.836	0.1312	1.108	0.2089
TE 03n	2.225	9.32	0.288	1.207	0.0905	2.330	0.302

NOTATION: λ = FREE SPACE WAVELENGTH OF CAVITY RESONANCE
λ_g = GUIDE WAVELENGTH
d = DIAMETER OF ORIFICE
$W_a = \frac{1}{Q_a}$ = CAVITY LOADING

NOTE: FOR FEED LIKE CASES 2 AND 3, BUT WITH WAVEGUIDE TERMINATED IN BOTH DIRECTIONS, DIVIDE W_a BY 2 cavity are given on Table 4. The assumptions are that the orifice is in a wall of negligible thickness, its diameter is small compared to the wavelength, it is not near any surface discontinuity, and the wave guide propagates only its principal (gravest) mode and is perfectly terminated. In some applications, the computed diameter is somewhat smaller than experiment shows to be correct.

Coupling by means of an electron beam can also be used, but, owing to transit time, a re-entrant-type cavity is usually used to keep down the distance the electrons must travel within the resonator.

BIBLIOGRAPHY

1. Hansen, W. W., and R. D. Richtmyer, On Resonators Suitable for Klystron Oscillators, *J. Applied Phys.*, Vol. 10, 189–199 (1939).
2. Hansen, W. W., A Type of Electrical Resonator, *J. Applied Phys.*, Vol. 9, 654–663 (1938).
3. König, H., The Laws of Similitude of the Electromagnetic Field, and Their Application to Cavity Resonators, *Hochf. tech u. Elek:akus*, Vol. 58, 174–180 (1941). Also *Wireless Engineer*, Vol. 19, 216–217, No. 1304 (1942).
4. Wheeler, H. A., Formulas for the Skin Effect, *Proc. I.R.E.*, Vol. 30, 412–424 (1942).
5. Barrow, W. L., and W. W. Mieher, Natural Oscillations of Electrical Cavity Resonators, *Proc. I.R.E.*, vol. 28, 184–191 (1940).
6. Chu, L. J., Electromagnetic Waves in Elliptical Hollow Pipes of Metal, *J. Applied Phys.*, 9, 583–591 (1938).
7. Borgnis, F., Elektromagnetische Eigenschwingungen dielektrischer Räume, *Ann. Physik*, Vol. 35, 359–384 (1939).
8. Borgnis, F., Die konzentrische Leitung als Resonator, *Hochf. tech u. Elek:akus*, Vol. 56, 47–54 (1940). Resonant modes and Q of the full coaxial resonator. For long abstract, see *Wireless Engineer*, Vol. 18, 23–25 (1941).
9. Wilson, I. G., C. W. Schramm, and J. P. Kinzer, High Q Resonant Cavities for Microwave Testing. *B.S.T.J.*, Vol. 25, 408–434 (1946).
10. Kinzer, J. P., and I. G. Wilson, End Plate and Side Wall Currents in Circular Cylinder Cavity Resonator. *B.S.T.J.* Vol. 26, 31–79 (1947).
11. Kinzer, J. P., and I. G. Wilson, Some Results on Cylindrical Cavity Resonators. *B.S.T.J.* Vol. 26, 410–445 (1947).

POWER SUPPLY

By J. E. Young

23. RECEIVER POWER SUPPLY

The first receivers to employ vacuum tubes used batteries for filament and plate power supplies, but tubes were soon developed which had cathodes suitable for heating by alternating current. At the same time rectifier-filter combinations were developed to supply plate voltage, and it was no longer necessary to depend on batteries, where a-c power was available. During the ensuing years the size of batteries has been reduced, and their shape has been adapted to receiver applications. At the same time their life has been considerably increased. Batteries are now used extensively in receivers in areas where no power is available and for portable receivers which are often designed so that either battery or a-c power may be used. Sealed storage batteries have also been developed, and receivers using them arranged so that, if the receiver is plugged in an a-c line, the batteries are automatically recharged.

FILAMENT POWER. The filament power for a-c operated receivers is usually obtained from one or more windings of the receiver power transformer. The design of power transformers is covered on pp. 6-26 to 6-30. A common winding is customarily employed to excite all cathodes; however, it sometimes happens that both filamentary and indirectly heated cathode tubes are used in a receiver, and bias is applied to the filamentary tubes in such a way that the filament winding of the transformer is maintained at a potential considerably above ground. It is possible, in this case, that the difference in potential between the cathode and heater of the indirectly heated tubes will exceed the value specified by the tube designer. Where this condition exists, or where, because of the circuits in which the cathodes of the tubes are connected, it may likewise be possible to exceed the rated potential difference between the cathode and the heater for which the tube was designed, it is necessary to provide separate filament windings, grouping the tubes on these windings in such a manner that excessive voltage strain will not exist between cathode and heater of any tube.

PLATE POWER. The voltage supply for the plate circuits of the receiving tubes in a modern receiver is usually obtained from a vacuum-tube rectifier. A high a-c voltage is applied to the tube, and the rectified output is passed through a low-pass filter to attenuate the a-c components. In Fig. 1 a typical B supply circuit is shown. With a condenser

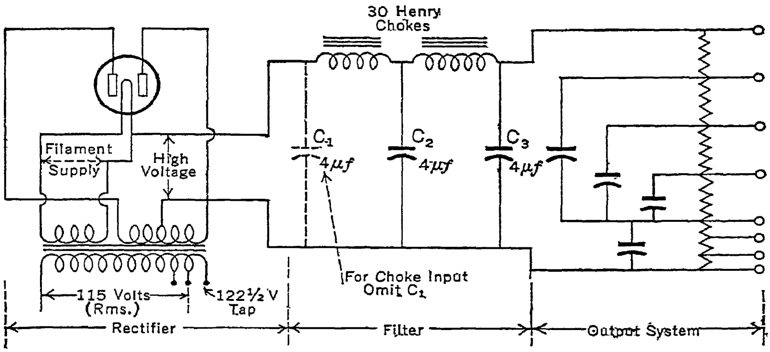


FIG. 1. Typical Receiver Power Supply

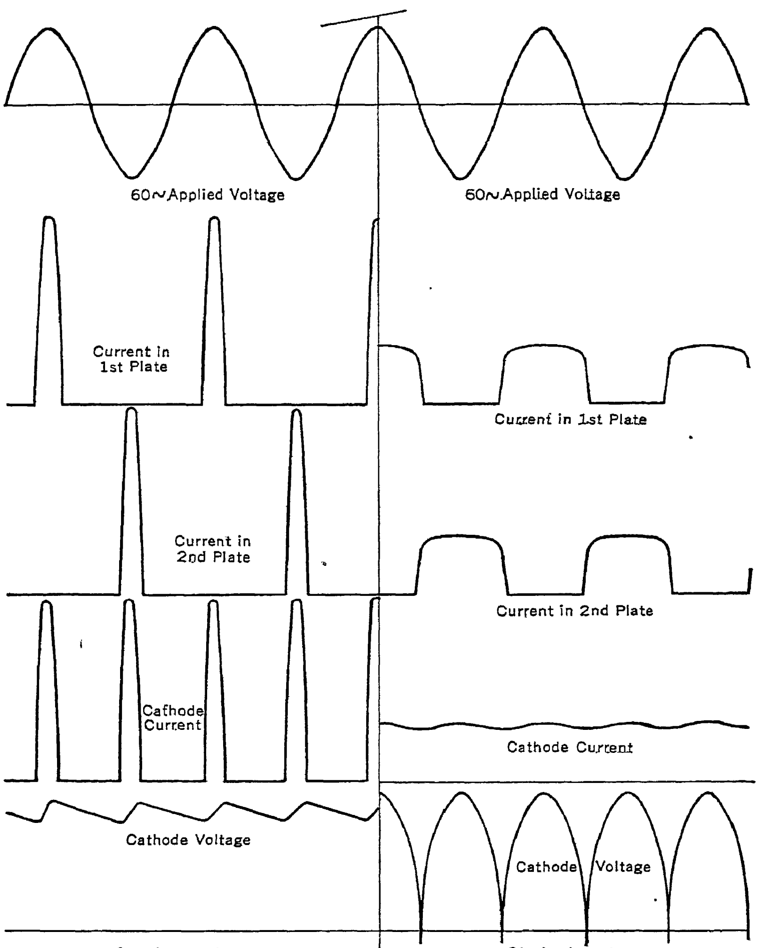
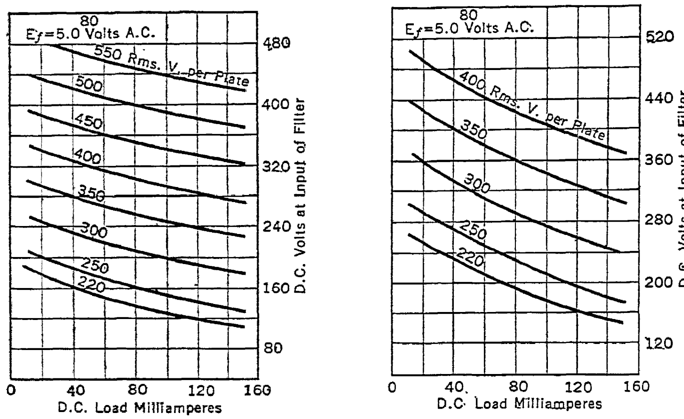


FIG. 2. Wave Form in Rectifier Circuits

input to the B filter (C_1 connected) the ratio of d-c output voltage to a-c input voltage is higher than with a choke input (C_1 disconnected). The choke input has the advantage of lower peak currents with less danger of damage to the tube under overload conditions. The voltage regulation with variable load is also better with choke input.

In Fig. 2 the form of the currents and voltages is given for various portions of these two circuits.

Figures 3*A* and 3*B* illustrate the relations between output voltage and load current for a typical full-wave rectifier for a choke input filter circuit and for a condenser input filter circuit.



A. Choke Input to Filter. *B*. Condenser Input to Filter.

FIG. 3. Output Characteristics of Rectifier Tubes

MERCURY-VAPOR RECTIFIERS. The efficiency of rectification is improved by the presence of mercury vapor in the tube. The mercury vapor reduces the voltage drop in the tube and improves the voltage regulation. This type of tube may produce r-f interference unless special precautions are taken, such as shielding of the rectifier tube and the use of r-f chokes or resistors in the rectifier plate leads.

B SUPPLY FILTER. The capacitor input filter is the most economical type if the transformer impedance and rectifier tube type are such that excessive charging current will not be experienced. In general, electrolytic capacitors are compact and inexpensive enough so that filters are usually designed with a rather large ratio of capacitance to inductance. It frequently happens that tubes, or the elements of tubes used for different purposes in the receiver, require different voltages. In this event it is economical to obtain additional filtering by using the voltage drop resistors and shunt capacitors as a resistance-capacitor filter. Since the capacitive reactance will generally be small compared to the resistance, the attenuation of this type of filter is inversely proportional to the product of the resistance and capacitance. If this is so, sufficient filtering is provided in the main filter to produce the required ripple attenuation for the tubes which are supplied directly from this filter.

It is possible to obtain increased attenuation by substituting a shunt resonant circuit for the inductive element of the filter; however, this practice is rarely followed, since the attenuation of ripple frequencies higher than the resonant frequency of the circuit is reduced and it is costly, in production, to maintain the values of inductance and capacitance closely enough to insure resonance at the correct frequency. The speaker field is frequently used as one of the chokes of the filter, and occasionally a hum-bucking coil is provided on the voice coil of the loudspeaker.

24. TRANSMITTER POWER SUPPLY

A-C POWER. A-c power is used for filament and plate supplies for transmitters almost universally. Filaments are usually heated by alternating current directly; rectifier-filter systems are provided for plate and bias supplies. Where alternating current is not available, motor generators or, for very low power transmitters, batteries are used.

FILAMENT POWER. Small transmitting tubes, unless used in service requiring quick-heating types, usually have indirectly heated cathodes. Unipotential cathodes are also required to permit operation at very high radio frequencies. Considerations affecting the design of filament power supplies for such tubes are covered in Section 6, article 14. Large tubes are usually directly heated and have either a tungsten filament or, where practicable, a thoria-coated tungsten filament. The emission efficiency of the latter is higher and less filament power is, therefore, required to achieve a given space current.

Filament power is usually derived from a 115- or 230-volt bus connected to the transmitter supply through a regulator, rheostat, or other means of holding the filament voltage within required limits (generally ± 5 per cent), and necessary control switches and protective devices. Step-down transformers convert the bus voltage to a value suitable for application to the filaments of the tubes. It should be noted that the life of tungsten-filament tubes is greatly affected by small changes in voltage. It is advisable, therefore, to provide means of adjusting the filament voltage of each such tube independently. The filament voltage should be no higher than is required to provide the necessary emission current.

FILAMENT STARTING. The cold resistance of transmitting-tube filaments is generally less than a tenth of the hot resistance. Filament-heating currents are usually large enough to cause severe mechanical stresses in the filaments and their supports. It is necessary, therefore, to limit the filament starting current to a safe value. This is determined by the designer of the tube and is frequently specified as one and one-half times the normal running current. Starting-current limitation is secured either by applying the filament voltages in steps, controlled so that the current reaches a steady value before the next voltage increment is applied, or by using a current-limiting reactor in the filament transformer primary circuit. The necessary reactance may be designed into the filament transformer itself, or a separate reactor may be used. This method of limiting filament current is preferable since the voltage increase across the filament terminals is smooth and is directly controlled by the filament resistance.

HUM DUE TO FILAMENT CURRENT. The electron stream emitted by the filament is affected by the magnetic field set up around the filament by the heating current. This effect is a periodic change in the space impedance between the filament and each of the other elements of the tube, at twice the frequency of the filament exciting current. The magnitude of this effect is a function of the design of the tube. The transmitter designer may minimize it either by the use of a tube having a multistrand filament connected to a two- or three-phase heating source, or by heating the filaments of tubes connected in pushpull or parallel from different phases of a two- or three-phase power supply. Tubes connected in pushpull and delivering power to the load at different parts of the audio-frequency cycle should have their filaments heated by currents which are in phase. The reduction of ripple obtained by multiphase filament connection will generally be of the order of 10 db for a two-phase filament connection and 14 db to 18 db for three-phase. The improvement obtained by the use of more than three phases is small, and the increased sensitivity of the ear to the higher resulting ripple frequency may make the hum more objectionable.

PLATE POWER. Except for emergency or mobile equipment, a-c power supplies are generally used. The a-c potential is rectified and filtered to obtain direct current having the requisite freedom from ripple.

TYPES OF RECTIFIERS. Selenium or copper oxide rectifiers are frequently used to obtain the relatively low voltages required for bias or for the plate circuits of low power tubes. For high-voltage plate supplies, rectifier tubes are employed. These may be of several types. The most common are vacuum tubes having a hot cathode emitter and a cold plate. To neutralize the space-charge drop, a small quantity of mercury or one of the inert gases is introduced into the tube. Other types of rectifier tubes use a pool of mercury as the cathode. Electron emission is obtained, either by maintaining an electron discharge from a hot spot on the surface of the pool by means of an auxiliary electrode excited by a separate power source or by discharging a heavy current through a concentrated point on the surface of the mercury by means of an electrode which just touches the surface of the mercury. The first of these two types of tubes has been called the pool-type, mercury-arc rectifier; the second is known as the Ignitron. The pool-type tube is available for single- or multiphase operation, whereas the Ignitron is commonly a single rectifier unit and a number are connected in groups for multiphase operation. The steel-tank mercury-arc rectifier has been frequently used for high-power applications such as railway power supplies. It has also been used to a considerable extent in Europe as a source of high voltage for radio transmitters. Because of its relatively high cost and the necessity for the provision of considerable auxiliary equipment, it has not been widely used in radio transmitters in the United States.

25. RECTIFIER CIRCUITS

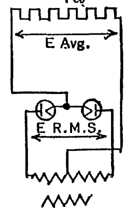
Figure 4 shows the common rectifier circuits for Kenotron and mercury-vapor tube type rectifiers with their voltage and current conditions. In calculating the d-c output voltage, from the "E average" given for the circuits shown, the voltage drop across the filter choke and the tube drop should be subtracted from E average to arrive at the d-c output voltage. The relations hold if a filter choke large enough to insure a constant load current is assumed. In practice, such a choke is usually required to obtain adequate ripple attenuation. Rectifier tubes and steel-tank rectifiers are rated on the basis of the maximum average plate current they can carry, the duration of the cycle over which the plate current flows, and the total voltage to which the tube is subjected during the period when it is not passing current. This voltage is known as the *peak inverse voltage*. In the hot-cathode, mercury-vapor tube, the safe peak plate current is also given, since, owing to the low internal tube drop, dangerously high currents may be passed through the tube, resulting in rapid deterioration of the filaments.

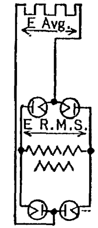
DOUBLE OUTPUT RECTIFIERS. In many transmitter applications it is desirable to provide two plate voltages, one for the low power stages and the other for the output amplifiers. If a single rectifier is to be used, the low voltages may be obtained by means of dropping resistors. However, if they are equal to or less than one-half the rectifier output voltage, one of the rectifier circuits which permits obtaining half-voltage is more economical. The single-phase bridge circuit or the three-phase full-wave circuit shown in Fig. 4, columns 2 and 5 respectively, may be so used. Half-voltage is obtained from a center tap on the secondary of the plate transformer in the single-phase full-wave circuit, or from the center of the wye-connected secondaries in the three-phase full-wave circuit. The additional transformer kilovolt-amperes required for the half-voltage loads can be computed by considering the equivalent rectifier circuits, which are the center-tapped circuit shown in column 1 for the single-phase rectifier or the three-phase half-wave circuit shown in column 3 for the three-phase rectifier. If the amplitude of the current to be supplied at half-voltage is an appreciable part of the total, a d-c component of current is produced in the secondaries of the high-voltage transformers, in the three-phase rectifier, which may seriously affect their operation. The effect of the d-c component may be eliminated by using a two-winding secondary for each phase of the wye, connected in broken star. It is generally desirable to use separate filters for the two voltage outputs in this type of rectifier; otherwise objectionable interaction may result, causing higher ripple voltage output than anticipated, or low-frequency feedback oscillation.

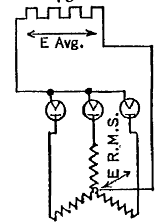
HIGH-VOLTAGE TRANSFORMERS. Transformers used in rectifier service are specially designed (see Section 6, article 14), since the insulation requirements and heating effects are quite different from those experienced in a-c circuit practice. Secondary winding insulation will depend on the type of rectifier circuit used. In the single-phase center-tapped secondary circuit shown in column 1, for instance, the center tap of the secondary winding is usually at substantially ground potential, whereas in the single-phase full-wave circuit shown in column 2 the midpoint of the secondary winding is at a potential equal to half the d-c voltage developed by the rectifier. Fault conditions will also seriously affect the insulation requirement of the transformer. In any of the circuits in which the center tap or midpoint of the wye of the transformer secondaries is connected to ground through the filter choke, a short-circuit fault will momentarily cause the full d-c voltage to be developed across the filter choke, raising the potential of the center point of the transformer secondaries to the full d-c voltage above ground.

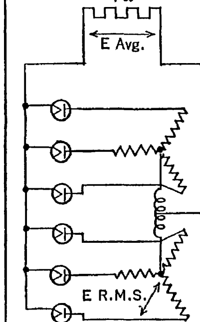
Transformers for three-phase rectifier circuits may be constructed as a single three-phase unit, or three separate single-phase transformers may be used. The first cost of the unit transformer will generally be lower, but having three single-phase transformers permits temporary operation in an open delta circuit if one transformer fails. Transformers may be obtained for rectifiers rated up to several hundred kilowatts and up to 10,000 to 15,000 volts d-c output, in either the dry or oil-filled types. Oil-filled transformers for any power are available, filled with Transil oil or one of the non-inflammable oils sold under various trade names, such as Pyranol or Dykanol. The non-inflammable oils require special transformer designs, since they will attack some of the insulation materials ordinarily used. In installations subject to the rules of the insurance underwriters, it is usually necessary that transformers filled with inflammable insulating oil be mounted in separate fireproof vaults provided with oil sumps and drains. The size of the transformer for which such protection must be provided is determined by its oil content and varies in the different states. Local codes should be checked.

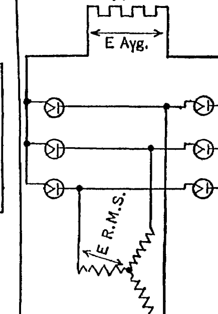
FILTER DESIGN. The design of the filter depends largely on the service for which the transmitter is to be used. Telegraph-transmitter filters are designed to the require-

4a


4b


4c


4d


4e


		4a	4b	4c	4d	4e
A	*E* Average	0.318 C	0.636 C	0.827 C	0.827 C	1.65 C
B	*E* R.M.S.	2.22 A	1.11 A	0.854 A	0.854 A	0.428 A
C	*E* Max	3.14 A	1.57 A	1.21 A	1.21 A	0.606 A
D	*E* Inverse	3.14 A	1.57 A	2.09 A	2.09 A	1.045 A
E	I_s R.M.S.	$0.707 I_{dc}$	I_{dc}	$0.577\ I_{dc}$	$0.289\ I_{dc}$	$0.816\ I_{dc}$
F	Secondary Kva.*	1.57	1.11	1.48	1.48	1.05
G	Primary Kva.*	1.11	1.11	1.21	1.05	1.05
H	Interphase Transformer Kva.	....	....	....	0.178	....
I	R.M.S. Ripple Voltage	48.3%	48.3%	18.3%	4.2%	4.2%
J	Peak Ripple Voltage	+36.3% −63.9%	+36.3% −63.9%	+36.3% −20.9%	+4.72% −9.30%	+4.72% −9.30%

* Primary and secondary kva are given in terms of d-c load kw.

FIG. 4. Rectifier Circuits

ment that the load current may vary at a rate corresponding to the telegraph characters, while the filters for telephone transmitters must, usually, be capable of supplying power at a very low audio frequency. It is usually necessary first to design the filter to secure the desired ratio between the load voltage and ripple voltage and then determine whether it fulfils other requirements.

Let F be the principal ripple frequency. Then F = supply frequency times number of rectifier phases. Let x_C = filter capacitive reactance; x_L = filter inductive reactance; C = filter capacitance in microfarads, and L = filter inductance in henrys.

Single stage:

$$\text{Per cent ripple} = \frac{mx_C}{x_L - x_C} \tag{1}$$

Double stage:

$$\text{Per cent ripple} = \frac{mx_C^2}{(x_L - x_C)^2} \tag{2}$$

where x_C and x_L are the capacitive and inductive reactances, respectively, of each of two similar stages; m = 70 for single-phase full-wave rectifier, 24 for a three-phase half-wave rectifier, and 5 for a three-phase full-wave rectifier.

Most Economical Filter Design. A single-stage filter is more economical than the double-stage type unless an unusually large reduction in ripple is desired, or the frequency is low, or low filter choke reactance is necessary, as for a telegraph transmitter.

Inductance-capacitance Ratio. The ratio between filter inductance and capacitance depends on a number of factors. If no other requirements are imposed on the filter than ripple reduction, the most economical ratio of L to C may readily be calculated. Ordinarily, however, the LC ratio is fixed by other considerations. If the rectifier tubes are worked near their current rating, the extra current flowing through the tube due to the impedance of the filter should be checked to determine whether the total tube current is excessive. This component of tube current may be calculated as follows:

$$I_p = \frac{\text{rms ripple voltage}}{x_L - x_C} \tag{3}$$

Filter for Telephone Transmitter. In most telephone transmitters the filter must supply an audio component of power, since the time lag through the filter and transformer reactances is too great to permit the audio component to be drawn directly from the transformer. The frequency and amplitude of the audio component depend on the type of modulation. The relations for the various systems are given below:

1. Class B Audio: Rectifier is required to supply an audio component having a peak value equal to the difference between the no-signal and maximum instantaneous signal plate currents of the class B stage at a rate corresponding to twice the lowest transmitted audio frequency.
2. Linear Amplifier and Grid-bias-modulated Amplifier: Rectifier is required to supply an audio component having a peak value equal to the unmodulated plate current multiplied by the modulation factor.
3. Constant-current: No audio component exists in the d-c power source.

It will be seen from the above considerations that the linear amplifier and the grid-bias-modulated amplifier impose the severest restrictions on filter design, while the constant-current system requires the filter to supply no audio-frequency power. The class B modulator requires some audio power, but the facts that the lowest audio frequency existing in the filter is twice the lowest modulating frequency and, further, that the class C modulated amplifier is usually supplied from the same source, drawing a steady current independent of the modulation frequency, make the filter design somewhat easier in this case.

Linear Amplifier. If we assume a linear amplifier, completely modulated, and it is desired to find a suitable filter combination, the following method may be used (if $r^2 > L/C$): Let F_0 = resonant frequency of L and C; F_d = frequency at which distortion begins; F = any audio frequency, to be investigated; K = ratio of load voltage at peak of audio cycle to unmodulated load voltage; K_d = above ratio at frequency F_d; r = load resistance; L = filter inductance; and C = filter capacitance. Then

$$K = \frac{a}{\sqrt{a^2 + b}} \tag{4}$$

where $a = F_0^2 - F^2$ and $b = \dfrac{4\pi^2 F^2 F_0^4 L^2}{r_2}$.

$$\frac{L}{r} = \frac{(F_0^2 - F^2)}{2\pi F F_0^2 K}\sqrt{1 - K^2} \tag{5}$$

$$F_d = F_0\sqrt{K_d + 1} \tag{6}$$

$$K_d = 1 - \frac{L}{r^2 C} \tag{7}$$

Sample Calculation. Assume that a linear amplifier requires an output of 10 kw at 15,000 volts. A filter consisting of a 15-henry inductance and 7.5-μf capacitor is provided. Determine the resonant frequency of the filter, the value of K at 30 cycles, the frequency at which distortion begins, and K at that frequency.

$$r = \frac{E^2}{W}\frac{225 \times 10^6}{10{,}000} = 22{,}500 \text{ ohms}$$

$$L = 15 \text{ henrys}$$

$$C = 7.5\ \mu\text{f} = 7.5 \times 10^{-6} \text{ farad}$$

$$F_0 = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{6.28\sqrt{15 \times 7.5 \times 10^{-6}}} = 15 \text{ cycles per second}$$

$$F = 30 \text{ cycles per second}$$

$$a = F_0^2 - F^2 = 225 - 900 = -675$$

$$b = \frac{4\pi^2 F^2 F_0^4 L^2}{r^2} = \frac{4 \times 9.85 \times 900 \times 50{,}600 \times 225}{506{,}000{,}000} = 797$$

$$K = \frac{-675}{\sqrt{465{,}000 + 797}} = -0.999$$

$$K_d = 1 - \frac{L}{r^2 C} = 1 - \frac{15}{506{,}000{,}000 \times 7.5 \times 10^{-6}} = 0.99605$$

$$F_d = 15\sqrt{0.99605 + 1} = 21.2 \text{ cycles per second}$$

Thus, it will be seen that, at 30 cycles, the reduction in rectifier voltage at the peak of the audio cycle is only 0.1 per cent, and distortion due to the filter circuit will not be encountered at any frequency above 21.2 cycles. At this frequency the reduction in voltage is 0.395 per cent.

Class B Modulator. The same method of calculation may be followed for a class B modulator, except that some allowance should be made for the steady d-c component of plate current supplying the no-signal plate current for the modulator and the modulated r-f amplifier. An approximate method is to find K as above and then to find the actual ratio, use the equation $K' = 2 - f(1 - K) - K$, where K' is the ratio of load voltage at the crest of an audio cycle to the unmodulated load voltage, corrected for the unmodulated component of plate current representing the sum of the class C amplifier plate current and the no-signal plate current of the modulator; f is the ratio of peak load current to load current without modulation; and K is calculated in the same manner as for a linear amplifier. It must also be remembered that only the double frequency component of the modulating frequency appears in the rectifier circuit, and so the modulating frequency should be doubled when it is used to evaluate F_0, F_d, and F.

Filter Chokes. Methods of calculation of the inductance required for the filter choke are covered in the section on filters. In addition to inductance, it is necessary to specify the voltage insulation or type of construction, the d-c current, and the a-c voltage. In normal operation the full ripple voltage of the rectifier will appear across the terminals of the choke. Its winding insulation must, therefore, be sufficient to withstand this voltage. In addition, a short circuit in the load will subject the choke to the full d-c rectifier output voltage, so that it must be designed to withstand this strain. The voltage insulation required between winding and core for normal operation will be lowest if the choke is connected between the rectifier and ground. This does not, of course, eliminate the necessity for the provision of adequate insulation to take care of load short-circuit.

Rectifier-tube Operation. The hot-cathode mercury-vapor tube is used in the majority of high-voltage rectifiers for radio transmitters. In addition to the limits of peak inverse voltage, and peak and average current set up in the rating of each of these types of tubes, it is also necessary that the condensed mercury temperature be maintained within specified limits. For operation of the tubes at their maximum rated peak inverse voltage, it is usually necessary to keep the condensed mercury temperature between the limits of 20 and 60 deg cent. For temperature ranges extending from 10 to 70 deg cent, the maximum peak inverse voltage is frequently halved. To control the condensed mercury temperature a jet of air is directed against a spot on the lower edge of the glass bulb, and the temperature of this air stream is occasionally controlled by means of auxiliary heaters. It is sometimes desirable to secure additional current capacity by connecting rectifier tubes in parallel. Because of the peculiar conduction characteristics of gases, unless such tubes are identical the one in which the gas is ionized first will conduct all the current and the other will not break down. This condition may be corrected by connecting a small center-tapped choke between the two tubes or by connecting a resistance in series with each tube so that sufficient potential is available to break down the second tube after the first one has started conducting.

TUBE HEATER DELAY. Most of the hot-cathode mercury-vapor tubes use highly efficient shielded cathods or filaments in order to reduce the filament power to a minimum. Such cathodes require some time to come up to their operating temperature, and it is usually advisable to provide a time-delay relay to prevent the accidental application of plate voltage before the cathode has reached its operating temperature. For the same reason, and to prevent the adherence of any particles of mercury to the anode or cathode, each new tube should be baked out thoroughly before plate potential is applied, and then reduced plate voltage should be applied, slowly working up to the normal operating plate voltage. Unless these precautions are followed, severe arc-backs may result and the tube will be permanently damaged.

TUBE-FAILURE PREDICTION. Mercury-vapor rectifier tubes almost always fail by arcing back, that is, becoming conductive to a voltage of either sign. This condition will occur, momentarily, and then clear itself; however, as the tube ages, it happens with increasing frequency, until it cannot be tolerated, and the tube must be replaced. Each arc-back short-circuits the plate transformer and usually trips the a-c overcurrent relays. If a bank of tubes is used in a multiphase rectifier it is difficult to determine by visual observation which tube has arced back. Devices which will register the flow of reverse current, such as polarized magnetic drops, are sometimes used as indicators. However, the short-circuit current is often so great that it will cause the indicators to drop on other tubes as well as on the defective one. It is possible to predict with fair accuracy the time when a tube may be expected to fail by making a routine check of the arc-drop voltage when the tube is carrying rated current. When, on successive readings, separated by perhaps 100 hours of normal operation, the arc drop is found to be rising rapidly, the tube will probably soon fail and should be removed from service. *These tests must, of course, be made by removing the tube from its operating position, and applying the necessary test voltage, which need not be greater than 100 volts.*

RECTIFIER CONTROL SYSTEMS. Since high-power rectifiers must usually be designed to have low regulation, a fault, in the form of either a short circuit in the load or an arc-back in a rectifier tube or tank, may result in dangerously high currents in the system. To minimize any trouble resulting from this source, a high-speed breaker should be provided in the power transformer primary. The breaker should be controlled by a-c overload relays in each phase of the primary and by a d-c overload relay in the output circuit. If a short circuit should occur in the transmitter, the energy stored in the filter will be dissipated in the fault even though the primary circuit is cleared instantly. For this reason it is advisable to incorporate some series resistance in the load circuit to aid in dissipating the filter energy. A resistor of 1 to 5 per cent of the load resistance can usually be added with no bad effects. It should be remembered, in designing such a resistor, that for a gassy tube or similar fault the load resistance is virtually zero, and all the rectifier voltage will, for an instant at least, appear across the protective resistor. This should have sufficient thermal capacity to dissipate several times the energy stored in the filter and should be insulated to carry the full rectifier voltage across its terminals. As a further protection against high voltages across the power transformer secondary in the event of an arc-back, it is advisable to connect a spark gap in series with a current-limiting resistor between each high-potential secondary terminal and ground. The gap may take the form of either a horn or sphere gap and should be set to break down at about 1.5 times the normal voltage.

BIBLIOGRAPHY

Armstrong, R. W., Polyphase Rectification Special Connections, *Proc. I.R.E.*, January 1931.
Lee, R., Rectifier Filter Circuits, *Electric J.*, Vol. 29, April, 1932.
Lee, R., Radio Telegraph Keying Transients, *Proc. I.R.E.*, Vol. 22 (February 1934).
Prince and Vogdes, *Mercury Arc Rectifiers and Circuits.* McGraw-Hill (1927).

RADIO RECEIVERS

By Vernon D. Landon

The functions of a radio receiver are to:

First: Select a desired signal from the heterogeneous signals picked up by the antenna.

Second: Amplify the radio-frequency signal selected.

Third: Detect the signal, thereby producing audio-frequency currents. (In the case of continuous wave code signals, it is necessary to heterodyne the signal with a local oscillator before detecting.)

Fourth: Amplify the audio-frequency signal.

Fifth: Reproduce the signal audibly by means of a loud speaker or headphones.

The parts of a receiver performing the above functions sometimes have overlapping duties. For example, the antenna circuit gives some amplification due to resonance and has some selectivity.

The simplest antenna coupling circuit is shown in Fig. 1*A*, with its equivalent. r_a and C_a are the effective resistance and capacitance of the antenna, and L_s and r_s are the inductance and resistance, of a variable inductor, in the receiver. E_a is the voltage induced in the antenna by the incoming signal. The step-up ratio of the circuit is defined as the ratio of E_s to E_a. At resonance (neglecting that component of E_s due to r_s, $E_s/E_a = j\omega L_s/r$, where $r = r_a + r_s$. The step-up at a frequency other than resonance is $E'_s/E'_a = j\omega L_s/z$, where $z = r_a + r_s + j(\omega L_s - 1/\omega C)$.

The ratio of the step-up at resonance to that at a frequency differing from resonance by a given amount is known as the selectance, or the discrimination ratio, for the given frequency difference. A curve of selectance vs. frequency difference is a selectivity curve.

To a rather close approximation the selectance is equal to $S = 1 + j4\pi f_d L/r$, where f_d measures the frequency difference from resonance. Since the only circuit constants in this expression are L and r, the figure L/r is said to determine the selectivity of the circuit. The selectivity is not changed by a change of carrier frequency if L/r is kept constant.

The circuit of Fig. 1*A* has several disadvantages. The step-up is high and reasonably constant over the tuning range, but the selectivity is poor, owing to the large antenna resistance in series with the tuned circuit. Also it is very difficult to incorporate in a unicontrol tuning system. The circuits of Figs. 1*B*, 1*C*, and 1*D* are more commonly used.

In 1*B* a tunable circuit is connected to the antenna through a small coupling condenser C_c. If C_c is quite small (as it is in practice) then the antenna resistance and capacitance may be neglected with only a slight error. By the use of Thévenin's theorem the circuit then reduces to that on the right of Fig. 1*B*. This is a simple series circuit. The step-up of such a series circuit, considered by itself, is nearly a constant over the tuning range. It is exactly constant if r_s is exactly proportional to the frequency. In practice r_s usually varies slightly more rapidly than the frequency. However, in this case the input voltage varies with frequency. As indicated in the diagram the effective driving voltage $E'_a = E_aC_c/(C_s + C_c)$. Since $C_s + C_c$ is the capacitance which produces resonance, then $C_s + C_c$ is inversely proportional to the square of the frequency. Hence E'_a is directly proportional (and the output voltage E_s is also roughly proportional) to the square of the frequency. This is the chief disadvantage of this circuit. Its advantages are its good selectivity and the ease with which it may be incorporated in a unicontrol tuning system.

In Fig. 1*C* the tunable circuit is connected to the antenna through a large inductance. If this coupling inductance were quite large the capacitance and resistance of the antenna could again be neglected. If the power factors of L_c and L_s are assumed to be equal a transformation involving Thévenin's theorem gives the equivalent circuit shown. Here, since L_s and L_c are constant, the input voltage E_a is constant. If all these assumptions are correct and r_s is proportional to the frequency, the output voltage E_s is also constant. In practice there are three effects combining to produce a marked drooping of the step-up at the high-frequency end of the tuning range. First, the inductance of L_c is usually not large enough to make the antenna capacitance negligible. Hence, the effective inductance of L_c and the antenna in series is less at low frequencies. This increases the low-frequency step-up. Second, the distributed capacitance of L_c increases its effective inductance most at high frequencies and lowers the high-frequency step-up. Third, the resistance of r_s usually varies faster than the first power of the frequency, lowering the high-frequency step-up.

An antenna circuit which is often used involves a combination of capacitative and inductive coupling in order to obtain a flat step-up characteristic. This circuit is shown in Fig. 1*D*, with its equivalent. It may consist of a tuned secondary, of the usual type, coupled to a primary of about eight times the secondary inductance. Loose inductive

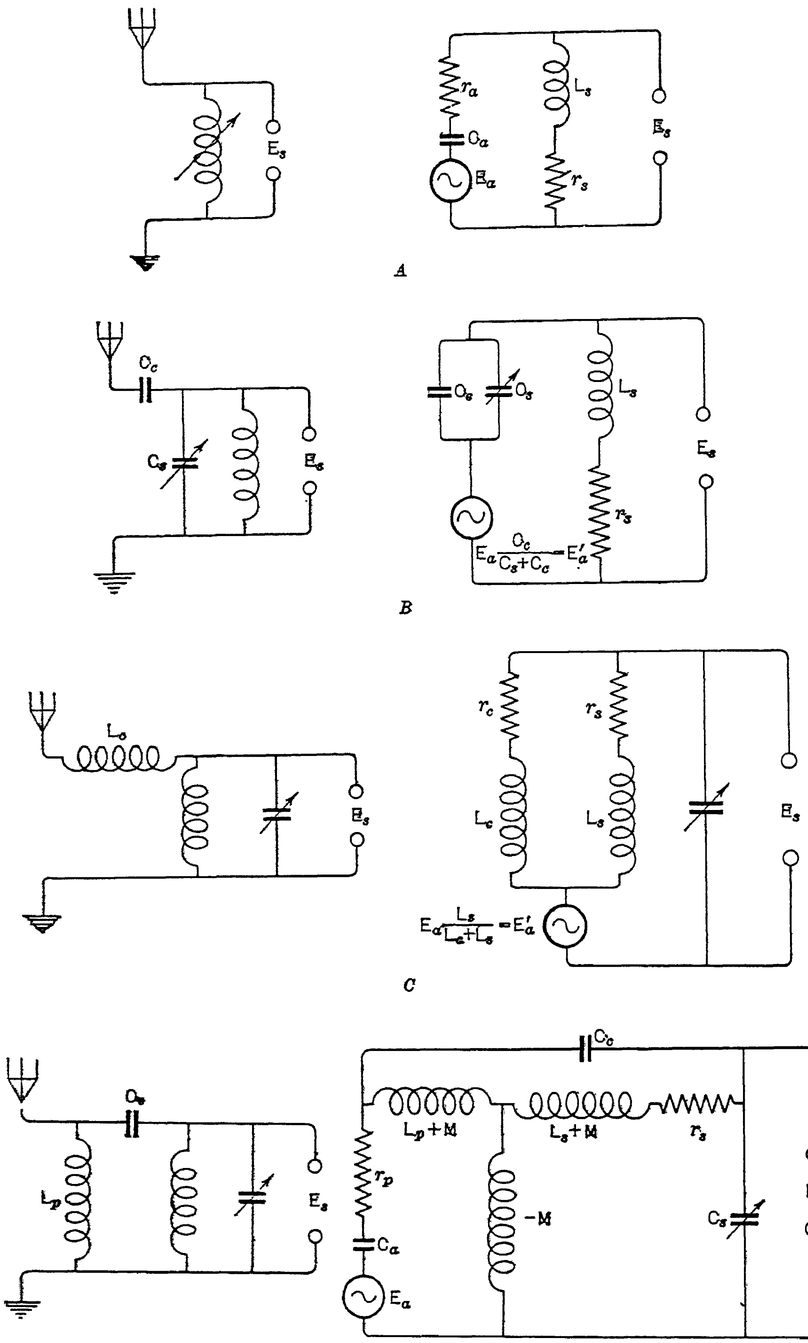


Fig. 1. Antenna Coupling Circuits

coupling is used. The capacitative coupling is adjusted to the value required to give the desired step-up at the high-frequency end of the tuning range. For uniform and maximum step-up it is essential that the inductive coupling have the proper phase, so that the capacitative coupling adds rather than subtracts. To obtain this condition the grid and antenna leads must emerge from the transformer with opposite directions of rotation. The step-up of such a transformer is practically constant over its tuning range. In a transformer for the broadcast band the value of the step-up usually lies between the limits of 3 and 10, depending on design constants. If tight coupling is used giving high step-up, the penalty is more detuning with changes in antenna constants. This results in poor tracking, with the other tuned circuits of the receiver, unless an antenna is used of the size for which the circuit was designed.

26. TYPES OF RECEIVERS

CRYSTAL DETECTOR RECEIVER. The simplest type of complete receiver is a crystal detector circuit such as shown in Fig. 2. The selectivity of this receiver is very slight, and amplification is lacking except for that due to resonance. Nevertheless local stations can be received.

For low-impedance crystals, the selectivity of the receiver can be improved by connecting the input to the crystal across a portion of the coil.

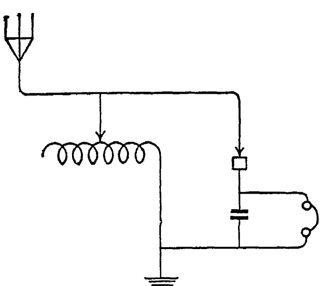


FIG. 2. Crystal Detector Receiver

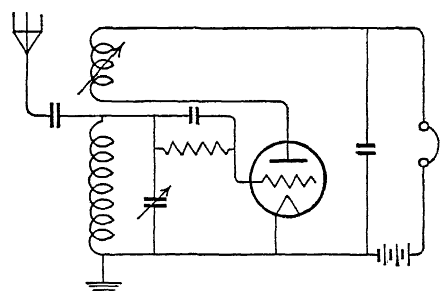


FIG. 3. Regenerative Receiver

REGENERATIVE RECEIVER. Figure 3 shows the circuit of a regenerative detector. The inductive coupling of the small coil in the plate circuit to the tuning inductance is adjustable. When the regenerative feedback is adjusted to a critical value, just less than that required to produce self-oscillation, a great amplification of signals results. The selectivity curve is much too sharp, resulting in a loss of sidebands, thus unduly impairing the fidelity of reproduction. The greatest objection to regenerative receivers is due to their ability to oscillate when the feedback is too great. This results in a radiated signal which produces very objectionable squeals, or beat notes, in near-by receivers. This type of receiver is now illegal. It can be made legal by placing a neutralized stage of r-f amplification ahead, so that it will not radiate when properly shielded.

Superregeneration of the Blocking Type. If the feedback in Fig. 3 is advanced well beyond the point of oscillation, the oscillations become self-modulated. This is due to a periodic blocking of the tube. The r-f voltage, rectified by the grid, produces a bias voltage across the grid leak sufficient to produce plate current cutoff, and oscillations die out. The charge on the grid condenser leaks off, the tube again starts to oscillate, and the blocking cycle repeats itself. The frequency of blocking depends partly on the tube but chiefly on the time constant of the grid leak and condenser. The higher the frequency of blocking, the greater the feedback required to produce the effect.

When the frequency of blocking is increased to about the limit of audibility, by decreasing the values of the grid leak and condenser, the circuit becomes an extremely sensitive receiver. It is even more sensitive than the regenerative receiver, and much less critical to adjust. The disadvantage is extremely broad tuning.

The sensitivity to weak signals is due to the fact that an oscillator cannot start oscillating in the absence of an impulse to start it. Weak impulses in the form of noise are always present. It is only necessary for the signal to exceed the random noise in order to control the oscillation. Since the peak amplitude of each block of oscillation is very closely the same regardless of signal amplitude, it is difficult to see how audio-frequency signals are produced. Probably the effect of the signal is to increase the frequency of blocking by causing oscillation to start sooner each blocking cycle. This decreases the plate current, since the tube is cut off a greater percentage of the time. Thus an amplitude-modulated signal produces audio-frequency currents, of the modulation frequency.

Superregeneration Employing One Tube Oscillating at Two Frequencies. Improved results over the above can be obtained with the circuit of Fig. 4. The tube of this circuit oscillates continuously at the quench frequency of about 15 kc, drawing grid current at one point in each cycle. When not drawing grid current the tube and circuit are in a suitable condition to oscillate at the frequency of reception. However, when only weak pulses or signals are present to trigger oscillation, the voltage does not have time to build up to full amplitude before it is quenched by grid current. Under these conditions the amplitude of the r-f voltage on the grid, when grid current starts, is proportional to the actuating signal amplitude.

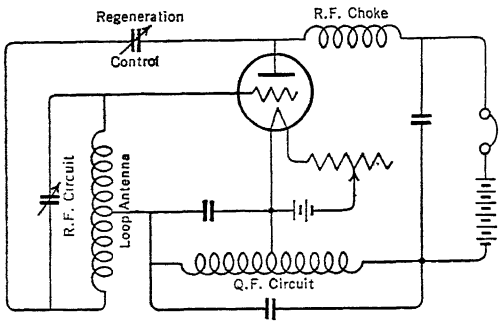


Fig. 4. One-tube Superregenerative Receiver

This circuit has a better signal-noise ratio than the blocking type, but it is almost equally broad in tuning. The chief application for these circuits is for reception at very high frequencies. For this use the broadness of tuning is frequently an advantage, helping to find and hold the signal.

THE TUNED RADIO-FREQUENCY RECEIVER. A tuned r-f receiver consists of several stages of tuned r-f amplification followed by a detector and audio amplifier.

Regeneration in Multistage Amplifiers. In a multistage tuned r-f amplifier, if capacitance exists between control grid and plate, regeneration will result. In fact, coupling of any sort between any two stages of an amplifier will result in regeneration, or oscillation, depending on the degree of coupling.

Resistance Stabilization. Since regeneration is equivalent to adding negative resistance, its effects may be largely counterbalanced, at a given frequency, by adding resistance to the input circuit. This added resistance should not be placed from grid to filament if it is desired to counteract regeneration over the whole tuning range with a fixed value of resistance. The regeneration is much more severe at the high-frequency end of the tuning range. Hence, it is desirable to place the added resistance in such a position that it will have its greatest effect at high frequencies. This is accomplished by placing the resistance in series with the grid.

The Tuned R-f Receiver with Resistance Stabilization. Figure 5 is the schematic circuit diagram for a complete battery-operated receiver, employing an untuned antenna

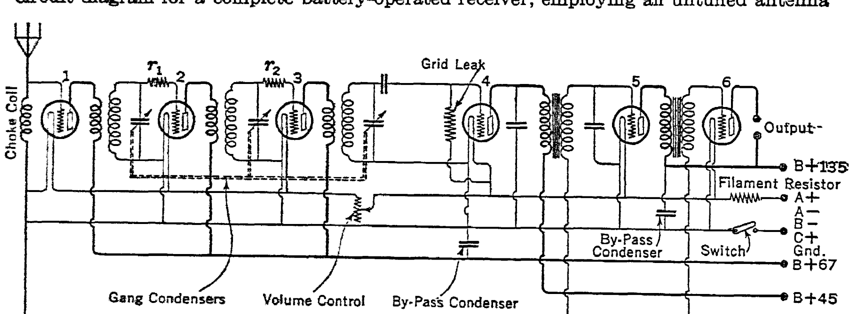


Fig. 5. Tuned R-f Receiver, Resistance Stabilized

circuit, and three r-f transformers, with resistance stabilization. Such a receiver would oscillate without the presence of resistors r_1 and r_2. When these resistors are given the proper value (usually about 800 ohms) the effects of regeneration may be approximately counterbalanced over the entire tuning range. In order to flatten the sensitivity curve the resistors are usually made large enough to overcompensate for the regeneration at high frequencies. For this reason the tuning is unduly broad at high frequencies, in this type of receiver.

The Neutralized Receiver. Another method of eliminating oscillation is by the use of neutralization. (See p. 7-29.) Receivers employing two and three stages of tuned amplification with capacitance neutralization were quite popular before the development of the screen-grid tube.

The Tuned R-f Receiver Employing Screen-grid Tubes. When screen-grid tubes are employed in a multistage tuned amplifier, regeneration of the type discussed above is not observed. The presence of the screen grid, between control grid and plate, reduces the grid plate capacitance to such a low value that regeneration is appreciable only when the stage gain is extremely high.

Nevertheless, it is necessary to take many other precautions to avoid oscillation if the overall gain is very great.

Other Sources of Regeneration. Coupling of any sort, between any two stages of the receiver, may give rise to serious regeneration. Coupling between adjacent stages is not as serious as between circuits which are one or more stages removed from each other. Capacitative, or inductive, coupling causes oscillation with equal facility because of the change in phase obtainable by tuning the intervening circuits. Incomplete shielding of grid and plate leads is one of the most prevalent sources of regeneration.

When the overall gain is high, the mutual inductance of the various sections of the gang tuning condenser becomes troublesome. Owing to the use of a common rotor shaft this coupling cannot be completely eliminated. It can be reduced to a satisfactorily low value by careful design. It is general practice to use several wiping contacts on the rotor shaft. The ground leads from the tuning inductances are brought separately to different terminals on the wiping contacts, to avoid the coupling of a common ground lead.

Objectionable coupling is often caused by the use of common voltage supplies for the cathode, screen, or plate circuits. This trouble can be eliminated by the use of small decoupling resistors in series with the voltage supply leads for each tube and with separate by-pass condensers.

THE SUPERHETERODYNE RECEIVER. The tuned r-f receiver requires extreme care to avoid oscillation, because of the high overall gain required at radio frequency. To avoid this the superheterodyne type of receiver, in which unduly high gain is not required at any frequency, is used. Part of the required amplification is obtained at the radio frequency and part at an intermediate frequency. This makes stabilization relatively easy.

The essential idea of the superheterodyne receiver is to amplify all signals at the same fixed frequency. The essential component parts are the preselector, the frequency converter, the intermediate-frequency amplifier, the audio amplifier, and the loud speaker. A typical receiver of this type is shown in Fig. 6.

The Preselector. The preselector consists of an antenna input circuit with, or without, one or more tuned r-f amplifier stages.

The operating characteristics of the preselector are identical with those of corresponding units in a tuned r-f receiver.

The preselector assists in producing discrimination against signals on adjacent frequencies, but the intermediate-frequency amplifier is so much more effective for this purpose that the use of the preselector is not warranted for this alone. Its essential function is the elimination of undesired responses at frequencies widely different from that of resonance.

The Frequency Converter. Frequency conversion is obtained by the use of the first detector and oscillator. In most receivers, the oscillator operates at a frequency higher than the signal frequency. The difference in frequency is the intermediate frequency. Voltage from the oscillator and from the signal is fed to the first detector, and the output is amplified in the i-f amplifier. Previous to 1933, a majority of receivers employed separate tubes for the oscillator and first detector, usually using a circuit in which the oscillator voltage was fed to the first detector cathode. A majority of modern receivers employ a single tube in which the two functions are combined. An example of this type of tube is the 2A7, used in the circuit of Fig. 6.

Combined First Detector and Oscillator. In the 2A7 tube the first two grids, adjacent to the cathode, comprise the oscillator elements. The voltage fluctuations of the first grid control the electron stream, not only to the oscillator plate (called second grid for convenience although it includes only two vertical rods), but also that to the remainder of the elements. The current arriving at the output plate of the tube consists of pulses, at the frequency of the oscillator. The amplitude of these pulses may be varied by the control grid. The remaining grid, or screen, acts as a shield between the oscillator elements and the control grid and output plate. It also screens the control grid from the output plate.

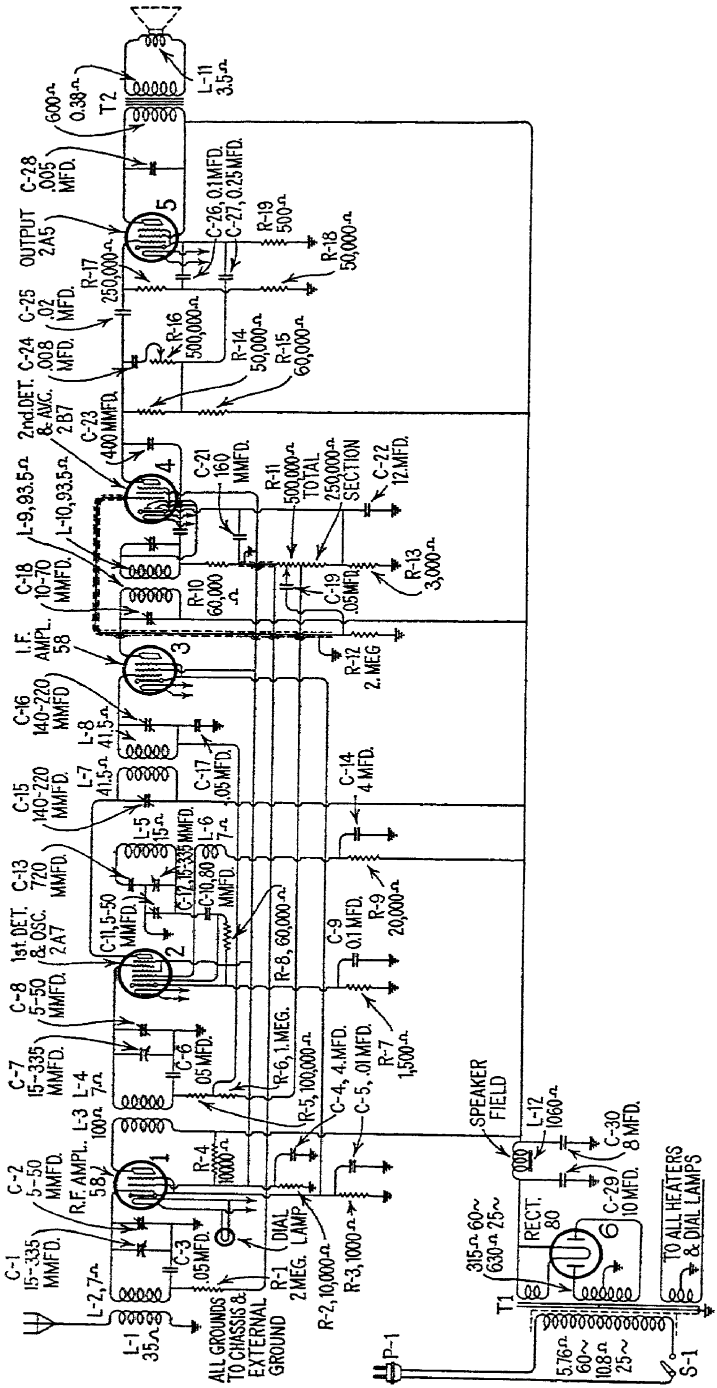


FIG. 6. Six-tube Superheterodyne

The component at the difference between oscillator and signal frequency corresponds to the intermediate frequency; it is selected and amplified by the i-f amplifier.

Tracking. One of the most important problems of superheterodyne design is tuning the oscillator and the preselectors with a gang tuning condenser. The problem is to maintain the oscillator at a uniformly higher frequency than the preselector, as the preselector is tuned over the frequency band. The frequency difference must remain equal to the intermediate frequency. Since the oscillator is operated at a higher frequency than the preselector, it has a tendency to change frequency too rapidly.

One method of correcting this is to specially shape the rotor plate of the tuning condenser which is used in the oscillator section of the gang tuning condenser. Another method is by means of fixed condensers in series and shunt with the oscillator tuning condenser. These reduce the rate of change of frequency. When the two auxiliary condensers and the oscillator inductance have the proper value, the oscillator frequency deviates only slightly from that desired, as shown in Fig. 7.

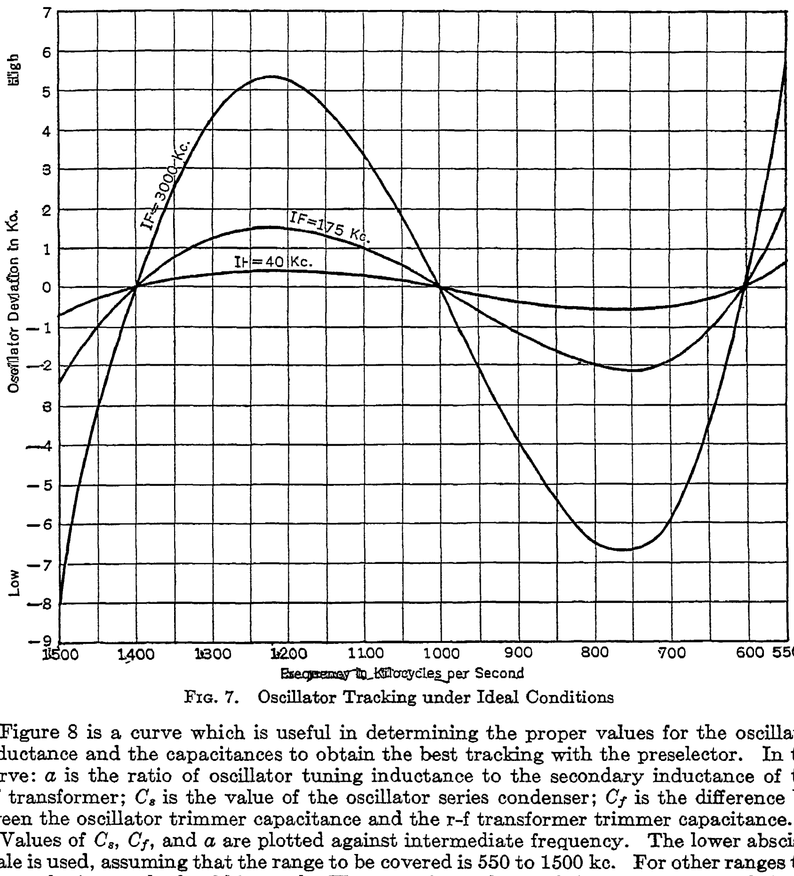


FIG. 7. Oscillator Tracking under Ideal Conditions

Figure 8 is a curve which is useful in determining the proper values for the oscillator inductance and the capacitances to obtain the best tracking with the preselector. In the curve: a is the ratio of oscillator tuning inductance to the secondary inductance of the r-f transformer; C_s is the value of the oscillator series condenser; C_f is the difference between the oscillator trimmer capacitance and the r-f transformer trimmer capacitance.

Values of C_s, C_f, and a are plotted against intermediate frequency. The lower abscissa scale is used, assuming that the range to be covered is 550 to 1500 kc. For other ranges the upper abscissa scale should be used. The curve is for the condition of 400 $\mu\mu$f total circuit capacitance at the low-frequency end of the range. If the maximum capacitance is changed, C_s and C_f change in the same ratio, while a remains unchanged.

The Intermediate-frequency Amplifier. The i-f amplifier consists of one or more stages of amplification following the first detector. The tuning is fixed at the intermediate frequency. Usually two coupled circuits are used in each i-f transformer. The frequency chosen usually lies between 100 and 500 kc per sec. High gain and good selectivity are easily obtained at these frequencies, particularly towards the lower of the two values.

The Diode Pentode Tube. In the circuit of Fig. 6, the fourth tube is a combination of a diode and a pentode in a common envelope. The diode and pentode together serve as detector and audio amplifier. This combination tube may replace two separate tubes in all circuits except those requiring different d-c cathode potentials.

Undesired Responses. Although the superheterodyne has many advantages, it is subject to a number of undesired responses and interfering beat notes which do not occur in a tuned r-f receiver. However, careful design minimizes these difficulties.

The Image Response. The most important undesired response in a superheterodyne is known as the "image." As explained above, the intermediate frequency is the difference between the signal frequency and the oscillator frequency. The oscillator is operated above the signal frequency. However, i-f signals are produced equally well by a beat

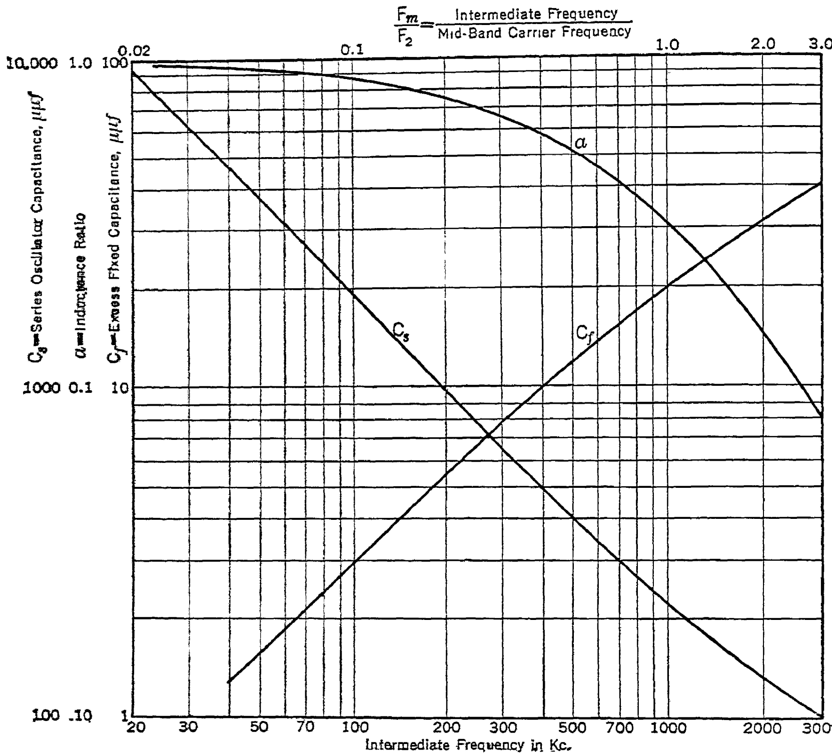


FIG. 8. Proper Values of Oscillator Inductance and Capacitances for Best Tracking with the Preselector

between the oscillator and a signal which is above the oscillator in frequency. The first detector is equally responsive to signals at either of these two frequencies. The only means of selecting the desired (lower frequency) of these two response points and attenuating the other is by means of the preselector. The higher the frequency of the intermediate amplifier, the greater the frequency separation of the desired signal and the image response; hence, the image response ratio is greater for a high intermediate frequency. With an intermediate frequency of 175 kc, the ratio of the sensitivity at the desired response to that at the undesired response can be made about 1000 at the high-frequency end of the broadcast range and about 10,000 at the low-frequency end. Although higher intermediate frequencies give higher image response ratios, other difficulties may develop from their use, as described below.

Harmonics of the Intermediate Frequency. Another source of difficulty which may be present in the superheterodyne is a beat note which occurs when reception is attempted at a frequency corresponding to a harmonic of the intermediate frequency. The reason for this beat note is that the second detector produces these i-f harmonics. If a very small amount of coupling exists between the second detector circuit and the antenna, or the r-f transformer, the i-f harmonic beats with the incoming signal producing a disagreeable squeal.

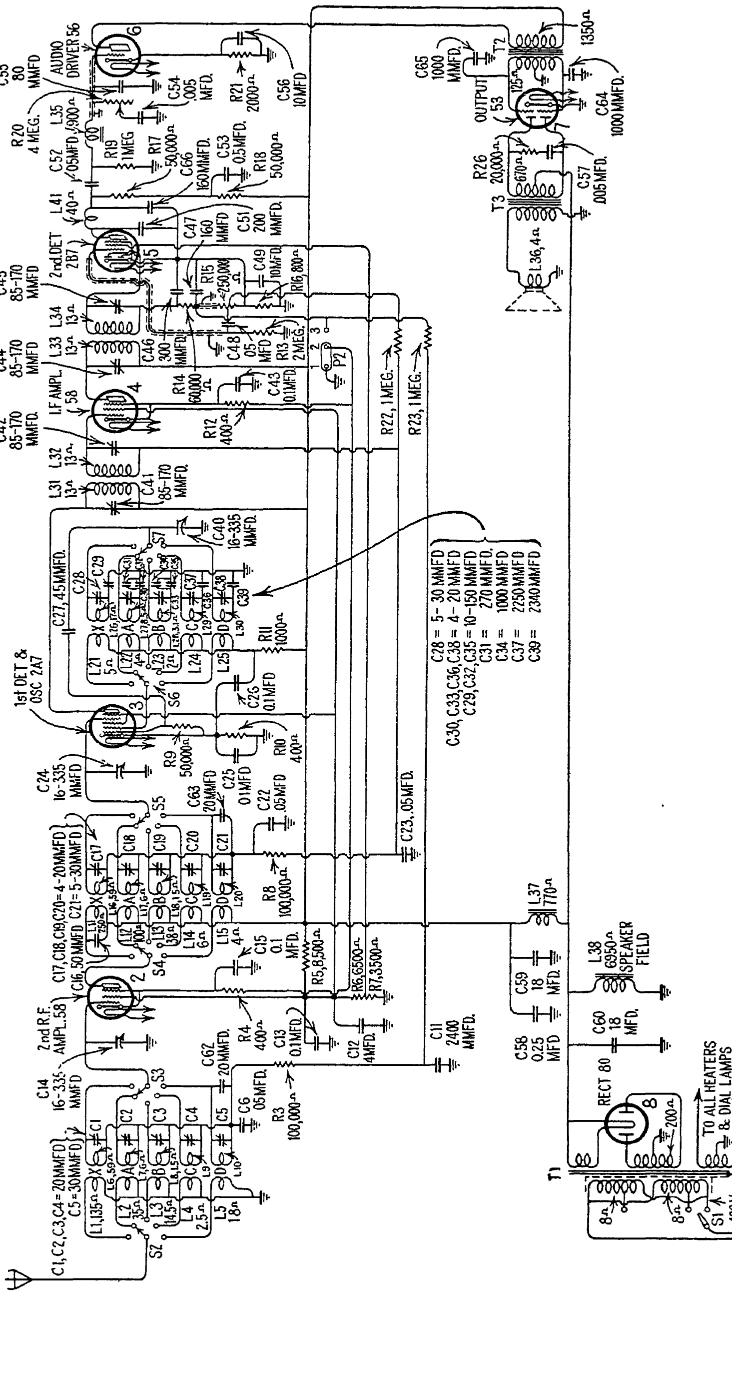


FIG. 9. All-wave Superheterodyne

The higher the order of the harmonic the less its amplitude in the detector circuit. Hence it is easier to suppress the beat note due to higher-order harmonics than that due to the second and third harmonics. The highest intermediate frequency which can be used if the third harmonic is to be kept outside of the broadcast frequency band is 175 kc. This accounts for the great popularity of this figure. When higher intermediate frequencies are used, such as 450 kc, the severity of the harmonics in the broadcast band is increased, but the number of interference points is reduced from five to two. By very careful shielding, the beat notes resulting from these harmonics may be almost completely eliminated. The greatest difficulty is obtained with the second harmonic.

Other Responses. The above are the most important of the undesired responses, but many other types occasionally give trouble. For example, two broadcasting stations may beat together to produce i-f signals independent of the local oscillator. Harmonics of the oscillator may beat with signals of various frequencies and with their harmonics, etc. A good preselector is the best insurance against all these.

ALL-WAVE RECEIVERS. The popularity of short waves increased very rapidly during 1933 and 1934. For this reason most of the commercial entertainment receivers now include provision for the reception of frequencies other than the standard broadcast band. Many of these receivers employ the name "all-wave," but this is a misnomer as none of these receivers cover the whole r-f spectrum. The circuit of a typical receiver of this kind is shown in Fig. 9. The circuit is entirely conventional, except that the preselector transformers and the oscillator transformers may be switched to any one of the five bands which it covers.

The r-f transformers for the low-frequency bands are purposely designed with restricted gain, so as to maintain approximately uniform sensitivity on all bands. This is necessary, because the gain obtainable in a single stage is limited to a low value at high frequencies.

RECEPTION OF CONTINUOUS WAVE CODE SIGNALS. In the reception of unmodulated code signals it is necessary to supply a local oscillator to produce an audible beat note with the incoming signal. In the regenerative detector circuit of Fig. 3 it is only necessary to advance the tickler to a point just beyond where oscillation starts in order to receive this type of signal. The regenerative detector may be preceded by one or more stages of tuned r-f amplification to increase the sensitivity.

The circuits of Figs. 5, 6, and 9 may be used to receive code signals by the addition of an external oscillator. In the tuned r-f receiver the oscillator must be tuned to beat with the signal directly. Hence it must be tunable over the receiving frequency range. With superheterodyne the local oscillator may beat with the signal at the intermediate frequency. Hence its tuning may be fixed. The oscillator should be coupled weakly to the detector input circuit.

TUNING INDICATORS. Because radio stations are not always modulating, it is somewhat advantageous to have a visual indication of resonance, rather than to depend on audio output as a check on the accuracy of tuning. There are many different methods for accomplishing this. One of the simplest is to place a milliammeter in the B supply lead to one of the r-f amplifier tubes which is subjected to automatic volume control. The stronger the signal, the higher the bias on this tube, and the lower its plate current. Hence the deflection of the needle downward from its peak value at no signal is a good indication of signal strength and of the accuracy of tuning.

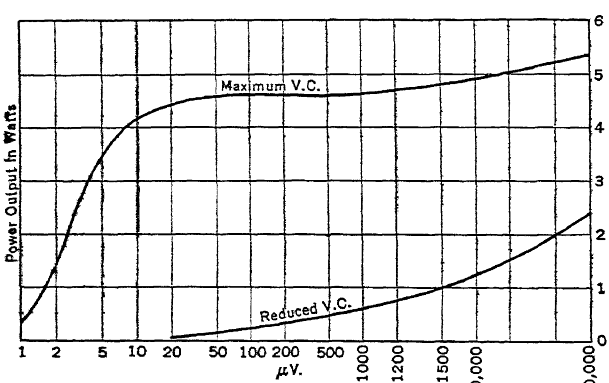


FIG. 10. Overload and AVC Curves

AUTOMATIC VOLUME CONTROL. The circuit connections for automatic volume control are shown in Fig. 6. The diode second detector develops a d-c voltage which is so connected as to increase the bias on the amplifier tubes with increased signal strength. The result of this connection is that strong signals produce only slightly greater audio response than weak ones. In Fig. 10 the audio output of a typical receiver is plotted against signal strength for maximum, and for a reduced manual volume control setting. A manual volume control, such as that shown in the diode circuit, is necessary to adjust the level of sound volume. After this adjustment is made signals come in at approximately the same volume. One of the important advantages of automatic volume control is a reduction of the effect of fading signals.

27. FIDELITY CHARACTERISTICS

The circuits affecting the fidelity characteristics are: the r-f amplifier, the automatic volume control circuit, the audio amplifier, the tone control, the output transformer, and the loudspeaker. The effect of the loudspeaker is not included in the curve of Fig. 11. Loudspeaker characteristics are discussed in Section 6.

The overall fidelity curve of Fig. 11 is the product of the fidelity curves of the component parts of the receiver.

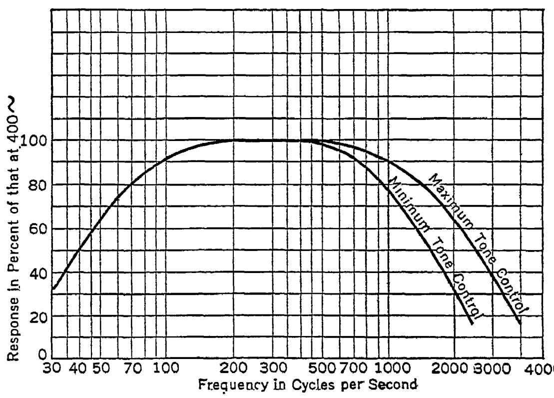


FIG. 11. Fidelity Curves Showing Effect of Tone Control

EFFECT OF R-F CIRCUITS ON FIDELITY. The r-f circuits affect the fidelity curve by cutting the high-frequency response. The modulation on a signal consists of continuous wave signals (called sidebands) on frequencies adjacent to the carrier frequency. These sidebands differ in frequency from the carrier by the value of the modulation frequency. The selectivity curve of the r-f and i-f amplifier shows quite appreciable decrease at only 2 or 3 kc from resonance. Hence the fidelity curve indicates this same decrease of the high audio frequencies.

THE EFFECT OF THE AUTOMATIC VOLUME CONTROL ON THE FIDELITY. The automatic volume control reduces the effects of fading, delivering to the detector a signal having only slight variations in amplitude, in spite of the wide fluctuations of signal amplitude on the antenna. In a similar manner the amplitude fluctuations of the signal, corresponding to low-frequency modulation, may be almost completely wiped out if the action of the automatic volume control is too fast. The action of this circuit may be slowed down by increasing the values of the resistors and by-pass condensers in the return leads of the tuning inductances. If the action is too slow the delay becomes noticeable to the ear. Shocks of static then blot out appreciable portions of the program, and the change in volume when tuning in stations may be noticeably slow. For this reason the time constant should be low enough so that there is a small but appreciable effect on the low-frequency portion of the fidelity curve.

THE RESISTANCE-COUPLED AUDIO AMPLIFIER. In Fig. 6 the diode section of the diode-pentode is resistance-coupled to the pentode section. Also, the pentode is resistance-coupled to the output tube. Neglecting the slight effect of the grid leak, the gain of a resistance-coupled amplifier is the ratio of the voltages applied to the grids of the preceding tube and the following tube. It is equal to $E_2/E_1 = \mu r/(r_p + r)$, where μ is the amplification factor of the tube, r is the load resistance, and r_p is the plate impedance of the tube. Or, if the plate impedance is very high, $E_2/E_1 = s_m r$, where s_m is the transconductance (or mutual conductance) of the tube at the operating voltages.

These formulas neglect shunt capacitance and the coupling capacitance. At high frequencies the shunt (plate-filament, grid-filament, and other) capacitances affect the result. The response is attenuated to 70 per cent of the mid-range value, at the frequency where the shunt capacitative reactance is equal to the effective resistance of r and r_p in parallel.

At low frequencies the coupling capacitance reduces the gain. The response drops to 70 per cent at the frequency where the reactance of the coupling capacitor is equal to the resistance of the grid leak. The effect of the diode resistance-coupling circuit may be calculated in a similar manner.

TONE CONTROL. The above paragraphs neglect the effect on the fidelity, of R-16 and C-24 (in Fig. 6), which constitute the tone control. At the maximum setting of the variable resistor the effect of these two units is very slight. However, when the control is turned back, high frequencies are progressively attenuated. A fidelity curve at maximum and minimum tone control setting is given in Fig. 11.

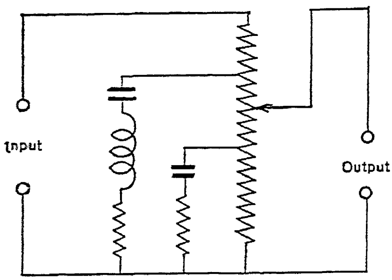


FIG. 12. Compensated Volume Control Circuit

The major use of the tone control is to improve the apparent signal to noise ratio. Noise is usually uniformly distributed over the audio-frequency spectrum, while the signal energy is chiefly contained in its lower frequency components. Hence when the tone control is turned back the signal is reduced by a smaller percentage than the noise.

The tone control may also sometimes be used to improve faulty fidelity in the transmitted signal.

THE EFFECT OF AUDIO TRANSFORMERS ON THE FIDELITY. Audio transformers of either the interstage or output type affect the fidelity by cutting both the high- and low-frequency response. However, the degree may be largely controlled by design. (See pp. 6-13 to 6-25.)

COMPENSATED VOLUME CONTROL. To the human ear the apparent loudness of sound at various frequencies changes at a different rate as the amplitude of the sound waves is changed. This makes it desirable to accentuate low frequencies and high frequencies when the volume control is turned down. A circuit employed to accomplish

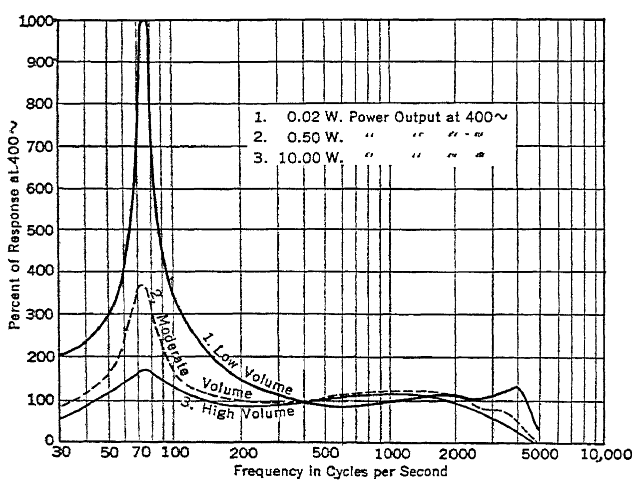


FIG. 13. Variation of Frequency Response with Setting of Compensated Volume Control

this is given in Fig. 12. Figure 13 gives the fidelity curve of a receiver incorporating this circuit. A receiver employing compensated volume control has a more natural sound at all volume levels.

NOISE SUPPRESSION. When automatic volume control is employed with a receiver of high sensitivity, the receiver automatically goes to full sensitivity when tuned between stations. The results are disagreeably strong reproduction of the static and general interference, which is always present in the background at any frequency. In order to make tuning more pleasant, circuits of various types have been developed for cutting off the audio amplifier in the absence of a signal carrier. One popular circuit for this purpose is shown in Fig. 14.

When no signal is present, no current flows in the diode circuit. Hence, no bias voltage is applied to the grid of V_2 and maximum plate current flows through r_2. The voltage developed across r_2 biases V_3 to cutoff so that it cannot amplify. The audio signals from interfering noises cannot pass this point. When an r-f signal is present, current flows in the diode circuit producing d-c and a-c voltages across the diode circuit resistor. The

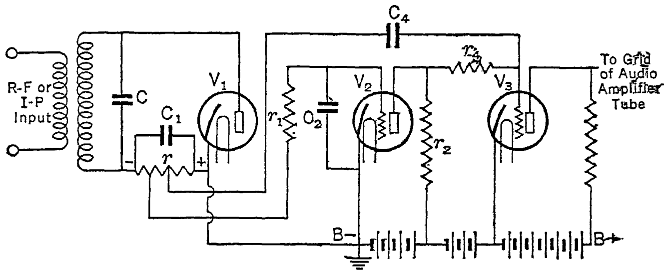


FIG. 14. Noise Suppressor Circuit

d-c voltage biases V_2 to cutoff. There is then no voltage drop across r_2, so V_3 operates with normal bias. V_3 then functions as an amplifier for the audio-frequency voltage applied to its grid through C_4.

28. RANDOM NOISE

Random noise (sometimes called fluctuation noise or Johnson noise) is a fundamental form of interference which prevents the satisfactory reception of signals below a certain level.

Random noise comes from two sources, thermal agitation in circuit resistances, and shot effect in vacuum tubes. Thermal-agitation noise comes from the random motion of electrons in a conductor due to its temperature. The open-circuit rms noise voltage across a resistor is

$$E_n = \sqrt{4KTr\,\Delta f} \tag{1}$$

where K = Boltzmann's constant = 1.37×10^{-23} joule per degree Kelvin.
T = absolute temperature in degrees Kelvin.
r = value of the resistance in ohms.
Δf = the effective noise bandwidth of the instrument used to measure the voltage.

Since the effects of shot noise are indistinguishable from those of thermal noise, it is customary to measure the shot noise of a vacuum tube in terms of the equivalent noise grid resistance. This is defined as the value of external grid resistance required to double the noise power output of the tube over that with the grid shorted. The value for a triode is approximately,

$$r_{\text{eq}} = \frac{3}{g_m}\,* \tag{2}$$

where g_m is the transconductance of the tube. The value for a pentode is approximately four times as high.*

Noise Factor. The "available power" of a signal generator is the power delivered to a load resistance under the condition of an impedance match. Thus the available noise power of a resistance considered as a noise generator is:

$$\frac{E_n^2}{4r} = KT\,\Delta f \tag{3}$$

If the resistance r is the impedance of a receiving antenna, and if no other noise sources existed, then a signal noise ratio of unity would be obtained when the signal power available from the antenna was $KT\,\Delta f$. This is the (unattainable) ideal which can never be

* B. J. Thompson, D. O. North, and W. A. Harris, Fluctuations in Space-charge-limited Currents at Moderately High Frequencies, *RCA Rev.*, Vol. IV, No. 3 (January 1940).

improved upon. Actually other noise sources always exist, so that the signal required for unity signal noise ratio is always greater than $KT\,\Delta f$. The ratio by which it is greater is called the noise factor. This ratio is usually expressed in decibels. Below 100 Mc receivers can be built that have a noise factor only a few decibels above thermal. Above 500 Mc, 10 db above thermal is considered good.

The test for "noise factor" is not yet accepted by the Institute of Radio Engineers as a standard test, but its use by the armed services during World War II became so widespread that its adoption as a standard seems inevitable.

The Nature of Random Noise. Random noise may be considered to be made up of an infinite number of sinusoidal components of different frequency. The amplitude of any single frequency component is infinitesimal, but in any finite bandwidth the rms voltage is proportional to the square root of the bandwidth though independent of the mean frequency.

Distribution of Amplitude. The actual voltage at any instant cannot be predicted, but an accurate statistical prediction can be made of the fraction of the time, taken over a long period, that the voltage will exceed any given value. This fraction of the time is identical with the probability that the given voltage V will be exceeded at a given instant and is equal to

$$P_V = \frac{1}{E\sqrt{2\pi}} \int_V^\alpha \exp\left(-\frac{V^2}{2E^2}\right) dV\,* \tag{4}$$

where E is the rms voltage of the noise.

Distribution of Envelope Amplitude vs. Time. If the bandwidth of the circuit passing the noise is small compared to the mean frequency, then the amplitude never changes abruptly from one cycle to the next. Thus, in a graph of the wave, if the peaks of adjacent cycles are connected with a smooth line the resulting line is the envelope of the wave and is a function varying much more slowly than the wave itself. The probability that the envelope will exceed a certain value A at any instant is

$$P_A = \exp\left(\frac{-A^2}{2E^2}\right)\,\dagger \tag{5}$$

The Values of Various Averages.† The average absolute value of the voltage is: $\bar{V} = 0.798E$.

The mean value of the envelope is: $\bar{A} = 1.252E$.

The root mean square deviation of the envelope from its mean value is:

$$A_r = 0.655E \tag{6}$$

The root mean square value of the envelope is: $A_{\text{rms}} = \sqrt{2}E$.

The most probable value of the envelope is: $A_p = \pm E$.

BIBLIOGRAPHY

Experimental Wireless and Wireless Engineering.
Proc. I.R.E.
Sturley, K. R., *Radio Receiver Design.* John Wiley (1943).
Terman, F. E., *Radio Engineering.* McGraw-Hill (1937).
Zepler, E. E., *The Technique of Radio Design.* John Wiley (1943).

RADIO TRANSMITTERS

By J. E. Young

A radio transmitter is defined as a device for producing r-f power for purposes of radio transmission. It also contains means of modulating or varying that r-f power, designated as the carrier wave, in correspondence to the intelligence it is desired to transmit. Television transmitters or others using the pulse technique, as well as those employing frequency modulation, are discussed in other sections. See Sections 8, 9, and 20. This section will be concerned with broadcast and communications transmitters employing amplitude modulation.

* V. D. Landon, The Distribution of Amplitude with Time in Fluctuation Noise, *Proc. I.R.E.*, Vol. 29, No. 2, pp. 50–55 (February 1941).

† K. A. Norton and V. D. Landon, Discussion on "The Distribution of Amplitude with Time in Fluctuation Noise," *Proc. I.R.E.*, Bol. 30, No. 9 (September 1942).

NATIONAL AND INTERNATIONAL REGULATIONS. Since radio communication, broadcast or point-to-point, involves transmission through a common medium, it has been necessary to set up national and international rules defining the frequencies, or channels, frequency tolerance, and type of emission for all radio stations. In addition, the Federal Communications Commission has set up national rules and standards governing frequency assignments, transmitter power, time when testing is permitted, specifications of performance regarding distortion, noise, and fidelity, grades of operators to be employed, etc., for all types of radio transmission. Other agencies, particularly the Radio Manufacturers Association, have been active in setting up recommended standards of performance and uniform methods of writing specifications and testing for various classes of transmitters.

RADIO TRANSMITTER—SCOPE. The transmitter is usually considered to consist of all audio equipment operating above standard telephone-practice levels, and all r-f equipment from the source of the r-f oscillations to the transmission line connecting the transmitter to the antenna. Within the equipment defined by these limits is contained sufficient audio-frequency amplification to raise the input signal to a level high enough to perform the function of modulation, r-f amplification and multiplication to raise the power level and frequency of the r-f oscillations to the output level, and the necessary power supplies for these circuits. The function and application of vacuum tubes to these elements of the radio transmitter are discussed in other articles of this section; only certain general aspects of design and performance will be covered here.

FREQUENCY CONTROL. The requirements of frequency stability are virtually always too severe to permit modulating (including c-w keying) an oscillator used as a direct source of antenna power. The usual practice is to use a low-power oscillator, separated from the modulated amplifier by a sufficient number of stages to prevent interaction caused by changes in impedance of the modulated stage resulting from the processes of modulation or other variations in load.

For transmitters operating on fixed frequencies, quartz crystals are commonly used as the frequency-determining element. A frequency stability of 10 parts per million is quite easily achieved with a quartz-crystal-controlled oscillator in which the temperature of the quartz plate is not permitted to vary more than ± 1 deg cent.

Some types of transmitters, particularly those used for military applications, operate on frequencies which may frequently be changed. For these applications quartz crystals are not practical and a master oscillator in which the tuned circuit is the frequency-determining element is employed. Such oscillators, having as much as 2 : 1 frequency range, may be designed to have a frequency stability better than 1 part in 10,000 for moderate variations of temperature, humidity, and power supply voltages.

OSCILLATOR POWER. It is possible to design crystal-controlled oscillators which will produce several hundred watts of output power; however, such oscillators are difficult to adjust and somewhat less stable than those of lower power output. It is better design practice to use a crystal oscillator which has only a few watts of output, followed by high-gain shielded-grid amplifiers, to achieve maximum frequency stability, and, in keyed-oscillator circuits, freedom from chirps and frequency creepage. This practice also permits the use of small crystals mounted in compact holders, since the crystal dissipates very little heat.

29. INTERMEDIATE-RADIO-FREQUENCY AMPLIFIERS

The intermediate-r-f amplifier stages perform the triple function of increasing the power level of the frequency-controlled oscillator, multiplying its frequency, if required, and acting as a buffer between the modulated amplifier and the oscillator. In the interest of simplicity, intermediate amplifier stages usually employ high-gain, multigrid tubes so that it is not uncommon to attain all three of these objectives with one or two intermediate amplifier stages. In high-level modulated and telegraph transmitters the intermediate amplifiers are operated class C. In low-level modulated transmitters the intermediate amplifier stages following the modulated stage must reproduce the audio-modulated envelope and must, therefore, be operated as class B r-f amplifiers. (See p. 7-22.)

INTERSTAGE COUPLING CIRCUITS. Amplitude-modulated transmitters are rarely used at frequencies above 40 Mc. Up to this frequency no special precautions are ordinarily required in interstage circuits. In general, it is desirable to make the circuits as simple as possible to avoid dangerous multiple resonances and parasitics. Typical coupling circuits for shielded-grid tubes are shown in Fig. 1. Note that the grid leak is not by-passed and that grid and plate feed circuits are dissimilar. It frequently happens that the input resistance of the driven amplifier is so low that the r-f choke in series with the grid leak may be eliminated without appreciably affecting the driver power output. This should always be done where calculations so indicate, since a possible parasitic circuit is thereby eliminated. For the same reason it is desirable to use the voltage obtained by the flow of rectified grid current through a resistor as grid bias rather than to obtain this voltage from a separate source. Protection against excessive plate dissipation, if the excitation fails, may be obtained by the use of sufficient cathode bias.

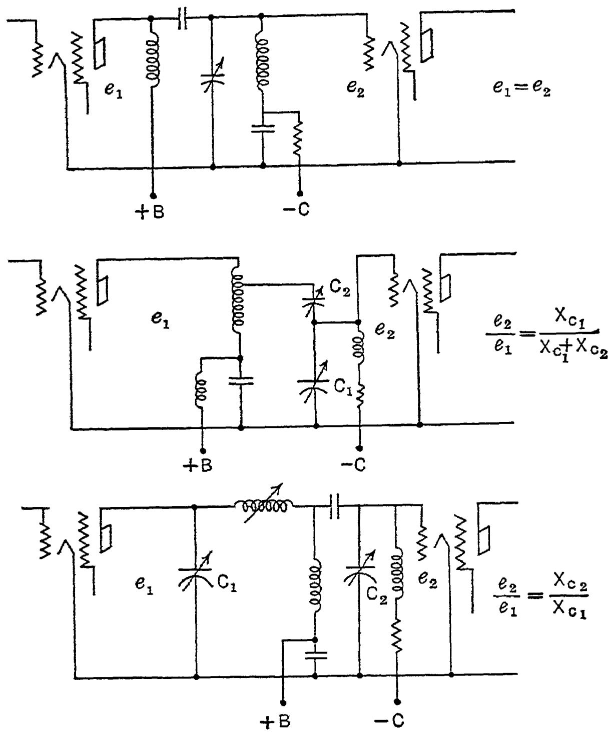


FIG. 1. Interstage Coupling Circuits

CIRCUIT Q, or KVA/KW RATIO. Unmodulated intermediate amplifier stages are not critical with respect to interstage coupling circuit Q, although it is desirable that the circuit, looking from the grid of the driven tube, be of high enough Q to insure sinusoidal wave shape in spite of the variable-impedance characteristic of the grid, as it is driven positive. Unfortunately, in the grounded filament circuits, the voltage relations are such that the driving tube puts energy into the coupling circuit almost 180° away from the time when it is absorbed by the grid of the driven amplifier. The energy storage must be sufficient to take care of this condition. The wave shape will be adequate if the circuit Q is greater than 10, unless the driven tube grid impedance at the peak of the positive grid swing is extraordinarily low. Unless the driving stage is modulated, even higher values of Q are useful and desirable, provided the concomittant coupling circuit losses are not thereby made excessive.

HARMONIC AMPLIFIERS. Because of special considerations, such as crystal activity, stability, etc., frequency multiplication is ordinarily employed where the carrier frequency is higher than 10 Mc. The necessary multiplication is accomplished in the intermediate amplifier stages. The multiplication through any one amplifier stage may be from twice to as high as five times. The plate efficiency of the amplifier tube is very nearly inversely proportional to the order of multiplication, if the most favorable conditions of loading and open angle of plate current are chosen for each multiplication. Because of the relatively low efficiency obtained when multiplying five times, this amount of multiplication is rarely used in any one stage. Because of the generally poorer plate-circuit efficiency, multiplication is usually accomplished at low power level, followed by amplifiers tuned to the output frequency, to boost the power level as required.

Optimum Angle of Current Flow. The term "angle of current flow" represents the portion of the grid voltage cycle during which plate current flows, expressed in degrees. The optimum angle of current flow depends on how nearly the plate-current plate-voltage characteristic conforms to the $^3/_2$ relation, and for a $^3/_2$ characteristic may be shown to be approximately 130° to develop maximum second-harmonic voltage in the plate circuit, 85° for the third, 65° for the fourth, and 50° for the fifth. The angle of current flow is given by the relation $\cos\frac{\theta}{2} = \frac{-E_c - E_0}{E_g}$, where θ is the angle of current flow in degrees of the excitation frequency, E_c is the grid bias, E_0 the projected cutoff bias, and E_g the grid excitation voltage. Negative bias voltages should be written in as negative numbers. Calculation will show that an open angle of 60° may be obtained with a bias voltage of not less than ten times cutoff bias. As the open angle is further restricted, the required bias voltage increases rapidly. It is not usually economical to use an open angle of plate current flow of less than 60° or to attempt to raise the frequency more than four times through a single amplifier stage. This does not hold true where the output power required is small, as in frequency-measurement work, where only milliwatts of power are usually required, and where the frequency may be multiplied many times.

30. POWER AMPLIFIERS

Power amplifiers may be either grid- or plate-circuit modulated. (See Modulators, pp. 7-73 and 7-74.) If the former, the power amplifier must be so operated as to reproduce accurately in its plate circuit an r-f envelope having the same wave shape as that of the excitation voltage. This may be accomplished by using either a linear amplifier or a

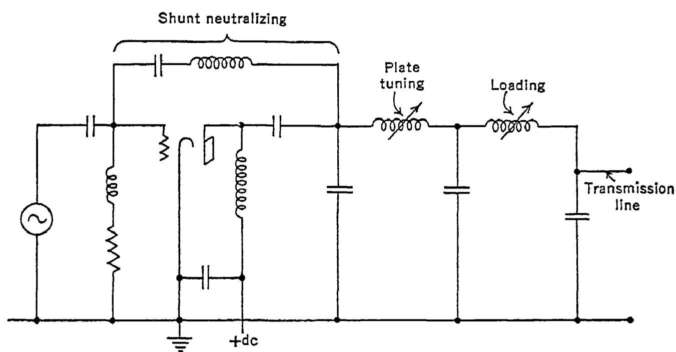


FIG. 2. Typical R-f Amplifier

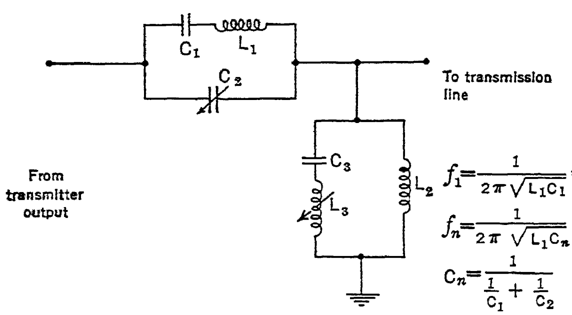
high-efficiency Doherty amplifier. The latter circuit utilizes linear amplifiers in pairs and provides for dynamic changes in drive and loading so that the average plate-circuit efficiency is several times that which may be obtained with a conventional amplifier. Two circuit paths are provided; in the first, the amplifier tube is adjusted to operate at the upper knee of its dynamic characteristic. Between its plate circuit and the load is connected a 90° phase-shifting network. The second tube is biased so that its plate current is almost zero under carrier conditions. As the drive increases during the positive excursions of the modulation cycle, the plate current of this tube, and its power output, increase in the same manner as an overbiased linear amplifier. The effective load resistance seen at the output of the 90° network connected in the plate circuit of the first tube increases as the second tube supplies power to the load. Because of the familiar impedance characteristic of the 90° network, this results in a reduction in resistance at the input terminals of this network. Therefore, the effective load resistance of the first amplifier is reduced and its power output is proportionately increased. Thus, at the peak of the positive excursion of the driving voltage, the power output of the first tube is doubled, because of its increased loading, and the output of the second tube is increased in the familiar fashion of a linear amplifier, so that it is also twice the carrier output. The total power output of the amplifier is thus four times the carrier level, fulfilling the condition for 100 per cent upward modulation. On the downward excursion of the modulation cycle the second amplifier is biased off completely, and the first amplifier functions as a conventional linear amplifier.

The envelope wave-form distortion resulting from the use of the Doherty amplifier is usually too great to be tolerable in high-fidelity systems. This defect may be remedied by applying overall feedback from the output terminals of the amplifier, through an r-f rectifier back into the audio-frequency amplifier circuits at some convenient point.

Amplifiers employing plate-circuit modulation are termed "high level." In practice the amplifier is operated class C, and the energy to perform the modulation function is provided by a class B modulator. (See Modulators, p. 7-74.) High-level amplifier output circuits must provide the necessary impedance transformation between the plate circuit of the tubes and the load and in addition must provide sufficient r-f harmonic attenuation to prevent excessive harmonic radiation. In practice, the Q of the output tank circuit is generally made quite low, from 4 to 8 for single-ended amplifiers, and additional harmonic attenuation is obtained, if necessary, by a low-pass filter inserted between the terminals of the amplifier and its load. A typical amplifier circuit employing shunt neutralization is shown in Fig. 2.

The high-level, modulated amplifier has the advantage that it is simple to adjust and uncritical in its operation. In addition, its quiescent (carrier level) efficiency is very high. Since the average percentage of modulation rarely exceeds 15 per cent in broadcast operation, the quiescent efficiency is a most important consideration in the economics of transmitter operation.

R-F HARMONIC RADIATION. Under the present conditions of crowded frequency assignments throughout the r-f spectrum, the first few multiples of the transmitter frequency—the frequencies that contain most of the harmonic energy—will often interfere



$$f_1 = \frac{1}{2\pi\sqrt{L_1C_1}} = \frac{1}{2\pi\sqrt{(L_2+L_3)\,C_3}}$$

$$f_n = \frac{1}{2\pi\sqrt{L_1C_n}} = \frac{1}{2\pi\sqrt{L_3C_3}}$$

$$C_n = \frac{1}{\frac{1}{C_1}+\frac{1}{C_2}}$$

f_1 = fundamental frequency

f_n = harmonic frequency

Fig. 3. Harmonic Attenuator

with some other radio service. Thus, because of a particular interference problem, one harmonic often must be reduced far more than the rest. The problem should first be analyzed, by making measurements to ascertain definitely that the interfering signal results from direct radiation from the station or its antenna. It frequently happens that a wire or metal surface located near the point of interference, or even the receiving antenna itself, is picking up energy from the transmitter and through the agency of an oxide-covered joint is itself producing the harmonic signal. This may be checked by field-intensity measurements made at a series of points on a line from the transmitter to the point of interference. Peaks in the harmonic signal intensity will be observed in the vicinity of such harmonic generators. The remedy is obvious but not always simple. Very often, however, a single guywire or other metal object having dimensions comparable to the wavelength in question will be found to be the source of the interfering signal.

The field-intensity meter may also be used to determine whether the interfering harmonic is radiated from the transmitter directly or from the transmitting antenna, if the two are sufficiently separated so that a directional fix may be obtained. If the radiation appears to be from the antenna, the signal energy may be carried by the transmission line from the transmitter to the antenna, or it may again result from rectification, either at an oxidized metal joint or at a tube rectifier, such as is used for remote antenna current reading or feedback. If the antenna rectifier is the offender the trouble may be cured by connecting a low-pass filter, designed to cut off just above the transmitter frequency, between the rectifier and the point where it picks up its energy from the transmitter output.

If the interfering signal emanates from the transmitter and is carried to the antenna by the transmission line, a trap, consisting of a circuit parallel-resonant to the harmonic, in series with the transmission line, and a circuit series-resonant to the harmonic in shunt with the line, will usually result in sufficient attenuation to eliminate the interference. Such a circuit may be designed so that its insertion will not affect the transmitter performance at its fundamental frequency, and it is preferably constructed so that the tuning of at least the shunt arm is variable, to permit exact adjustment to the harmonic frequency. Figure 3 shows one typical circuit.

If a balanced transmission line is used, the trap circuits should be symmetrical, with their center taps grounded to prevent in-phase transmission of the harmonic along the transmission line wires.

If the directional fix indicates that the transmitter itself is radiating the interfering signal, it may be necessary to resort to complete shielding to eliminate the trouble. More frequently, sufficient improvement will be obtained by other minor design changes, such as better grounding, more direct return of plate tank lead to filament of power amplifier or elimination of stray capacitance coupling between plate and output connection.

NEGATIVE FEEDBACK. The distortion, noise, and frequency characteristics of a telephone transmitter may be considerably improved by the application of negative feedback over all or part of the system. Two general methods of use of negative feedback are common. In transmitters using low-level modulation, followed by linear amplifiers, feedback voltage is derived from an r-f rectifier coupled to the output of the power amplifier. This signal, of the proper amplitude to secure the desired feedback, is introduced into one of the audio-frequency amplifier stages. Since the r-f rectifier is in the β loop of the feedback circuit, noise or distortion generated in it will appear in the ouput of the transmitter. It must, therefore, be carefully designed to reproduce accurately the audio envelope of the output r-f signal.

One difficulty frequently encountered in the application of this type of feedback to broadcast transmitters is that the transmitter antenna also functions as a receiving antenna for signals from other broadcast stations. Sufficient voltage may be developed in the circuit to which the feedback rectifier is connected by another broadcasting station, located in the immediate vicinity, to generate a series of cross-modulation products in the rectifier. Any of these new frequencies lying within the pass band of the α and β loops of the audio system will produce sidebands which will be radiated by the antenna and may cause serious interference. Such effects may be greatly reduced by installing a trap tuned to the frequency of the other station in the antenna system beyond the pick-up point for the feedback rectifier. Care must be used in designing such traps that they do not cause sufficient phase shift to alter the phase-shift attenuation characteristic of the transmitter seriously or instability may result.

The negative feedback will correct for distortion and noise, but only so long as there is reserve capability in the system to effect the correction. For example, it cannot correct for the distortion arising from overmodulation and, in fact, by introducing high-order harmonics back into the circuit may actually increase the distortion if the system capabilities are exceeded.

Since it is possible to keep the distortion of high-level modulated amplifiers down to a fraction of a per cent by the methods discussed under Modulators, p. 7-74, negative feedback is usually applied over the audio system, only, in this type of transmitter. This eliminates the need for the r-f rectifier and largely eliminates the effect of the output amplifier tank and load circuits on the performance of the feedback loop. The feedback voltage is usually obtained from a voltage divider, connected between the plate of each modulator tube and ground. The divider consists of capacitors, shunted by resistors. These elements are proportioned so that their impedances are equal at a frequency of approximately 100 cycles per second. The advantage of this type of divider over simple resistances is mainly that the operation of the circuit is unaffected by the capacitance of the leads connecting the network to the low-level audio-amplifier stage. Some improvement in the phase characteristic is also secured.

The use of feedback permits high-efficiency linear amplifiers to be used in services such as broadcasting, where low distortion and low noise level are essential. In high-level transmitters, feedback, applied over the audio system, including the modulator, effects a large reduction in noise and distortion and reduces performance variations due to non-uniformity of parts and tubes. It also improves the modulator efficiency, permitting the no-signal plate current to be reduced substantially to zero.

31. AUDIO AMPLIFIERS

Broadcast transmitters are, by industry agreement, provided with enough audio gain to produce 100 per cent sine-wave modulation with an audio input level of +10 db (1 milliwatt reference level) at an impedance of 600/150 ohms. Special-purpose transmitters may include more gain, although it is usually advisable to limit the gain so an input signal of at least −20 db is required. If enough gain is to be included in a transmitter to accept a lower signal level, special shielding and filtering precautions are advisable to prevent objectionable feedback.

Limiting amplifiers are generally used. A limiting stage may be designed into the audio system of the transmitter, or an external line amplifier of the limiting-compressing type may be used. Even in the latter case it is advisable to design one of the transmitter audio stages so that positive modulation greater than 125 per cent is impossible, regardless of the input signal level. Owing to switching transients or accidents, input levels 10 to 20 db above 100 per cent modulation are bound to occur, and, unless they are prevented from reaching the high-power amplifier by an absolute safeguard (such as driving the plate current of an audio stage to cutoff), serious damage may result.

Many special features are incorporated in audio systems to meet the requirements of particular services. Some of these are: automatic gain control, for circuits subject to wide variations in level; pre-emphasis of high audio frequencies, used on conjunction with complementary de-emphasis in the receiver to reduce noise; and scramblers which invert or otherwise distort speech, making it unintelligible unless a receiver provided with a complementary restorer is used.

32. TELEGRAPH TRANSMITTERS

For point-to-point transmission, telegraph code signals are often employed, since it is generally possible to secure 100 per cent transmission with a much weaker signal when telegraph is used instead of voice. Transmission may be accomplished either by tone modulation, which directly replaces the voice transmission, by keying the carrier wave directly (usually by overbiasing one of the intermediate amplifiers), or by shifting the frequency of the carrier. The last method may be accomplished by shifting the carrier a few hundred cycles back and forth at an audio rate and keying the audio signal or by shifting the carrier between two fixed frequencies, several hundred cycles apart, one corresponding to the "mark" and the other the "space." Both these systems are akin to frequency modulation and provide some advantage in noise suppression and diversity effect without the complication of diversity receivers.

33. INSTALLATION OF RADIO TRANSMITTERS

TRANSMITTER TESTING. A great deal of test equipment is now available, making it possible to examine conveniently all phases of the performance of a transmitter. Tests generally performed are as follows:

1. Circuit check—by means of click meter, ohmmeter, etc., to be sure that wiring has been done in conformance with wiring diagrams.
2. Control circuit check. Without tubes in any circuits, check the functioning of line switches, starters, time delay relays, overload relays, interlocks, and circuit breakers. No-load voltages of filament transformers may be checked at this time.
3. Transformer voltages. With filament circuits energized and tubes in place, check filament voltages, transformer tap settings, range of primary voltage control, functioning of air blower and/or water-cooling system.
4. Operational check. With all circuits tuned and functioning normally, check voltages and currents on all elements of all tubes. Check for instability and parasitic oscillation.
5. Modulation characteristics. Under normal operating conditions, check audio harmonic distortion at various modulation levels and at modulating frequencies throughout the normal frequency-transmission range. The same test equipment is required for measurement of the frequency characteristic, and residual noise level, so these tests should be performed at the same time. Figure 4 shows the equipment set-up required for this test.
6. Power output. This test should be performed using loads representing the high- and low-limit load impedances which the transmitter may be required to feed. To obtain accurate power checks, the load resistance should be accurately measured and the load

current determined by means of an accurate r-f ammeter; or the calorimetric method of power measurement may be used, in which the increase in heat content of water used to cool the dummy load is measured. The temperature rise of the water in degrees centigrade multiplied by the number of gallons flowing per minute multiplied by the constant 264 will be equal to the number of watts of power dissipated in the load.

7. Heat run. The transmitter should be run under conditions of normal operation for a period long enough for all components to reach approximately constant temperature. The procedure set up by the RMA committee on amplitude-modulated broadcast transmitters is a good example of how such a test should be run:

"The transmitter shall be operated (at rated power output) . . . long enough for all components to attain temperature stability; that is, until the hourly increment does not exceed 5 per cent of the total change under the following conditions: . . . The carrier

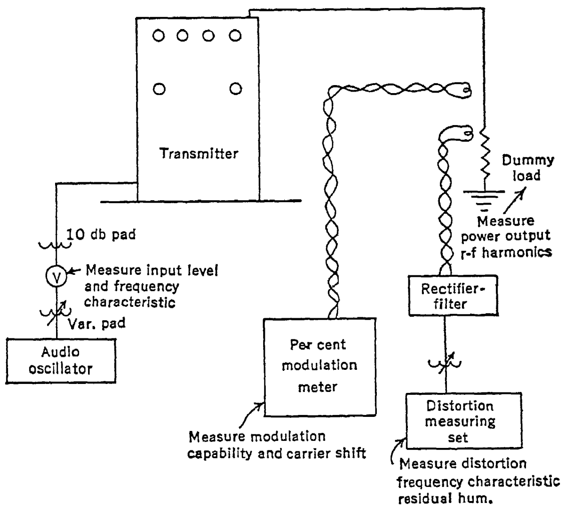


FIG. 4. Modulation Characteristics Measurements

should be continuously modulated by a 1000-cycle sine wave. The audio input level to the transmitter shall be approximately 10 db below that corresponding to 100 per cent modulation for not more than 98 per cent of the duration of the run and not less than 10 db above that corresponding to 100 per cent modulation for the remainder of the time." *

Temperature rise measurements of motor or generator windings, transformer windings, etc., should be made by the hot-cold resistance method, in which the resistance and temperature of the winding are measured after the unit has been out of operation long enough to reach a uniform, stable temperature. The resistance at the end of the heat run is then measured, and the rise in temperature is computed by the formula:

$$T = \frac{234R_2}{R_1} - \frac{234R_2T_1}{R_1} - T_2$$

T_1 = temperature at which cold resistance was measured.
T_2 = ambient temperature at end of heat run.
R_1 = resistance at T_1.
R_2 = resistance at end of heat run.
T = rise of winding temperature above final ambient.

8. Harmonic radiation. If the transmitter is tested using a dummy load, it is possible to make a rough check on harmonic intensity by coupling a field-intensity meter to the load through an attenuator, shielding the field-intensity meter so that all its pick-up is derived from the load. However, since dummy loads rarely ever even approximate the impedance characteristic of an actual radiating antenna system, such tests are of little quantitative value, and more satisfactory results will be secured if these measurements can be made on an actual field installation.

* See Radio Manufacturers Association, Engineering Department, Standards Proposal 172, Electrical Performance Standards for Standard Broadcast Transmitters.

9. Incidental phase modulation. Instruments which will measure the degree of phase modulation at frequencies below 40 Mc are seldom available. The development of commercial equipment operating in the 88–108 Mc band for the purpose of measuring the percentage of modulation of f-m broadcast stations, however, makes a handy tool available for this purpose. The signal to be measured is multiplied up to a frequency within the range of the meter by low-power harmonic amplifiers, and, since the meter is calibrated in terms of a 75-kc swing, the actual swing may be measured, and then converted into degrees of phase modulation, and divided by the multiplication ratio to determine the actual amount of phase modulation at the transmitter carrier frequency.

10. Telegraph transmitter tests. In addition to the above tests, measurements are made on the character of the keyed pulses of telegraph transmitters. The shape of the keying wave, build-up, transients, breaks, etc., may be observed or photographed on the face of a cathode-ray tube which has one pair of plates connected to a sample of the r-f output of the transmitter, and the other pair excited by a sweep voltage which is preferably adjusted to synchronize the picture of the keyed pulse on the screen. The mark-to-space ratio may be calculated by measuring the d-c current output of a rectifier energized from the r-f output of the transmitter. The percentage of mark-to-carrier is given by the ratio of the d-c current during keying to the current with key closed continuously. If the former is designated I_M, and the latter I_C, the mark-to-space ratio is

$$\frac{I_M}{I_C - I_M}$$

LOW-POWER TRANSMITTERS. Modern transmitters having a power output of 1 kw or less are usually designed as completely self-contained units. It is, accordingly, only necessary to provide power line, audio input, antenna or transmission line, and any necessary external monitor connections. Such transmitters are usually designed to operate from either a 115-volt or a 230-volt single-phase a-c source. Power and audio circuits may be conveniently brought into the transmitter through either conduits or trenches; trenches permit somewhat greater flexibility for future change or modification. External wiring and switching should be installed in accordance with fire insurance underwriters' specifications. Ground wires should be connected by as short a run as possible to a low-resistance ground, either the antenna ground system if any portion of it is installed near the transmitter or a separate ground consisting of copper plates buried in moist soil. Antenna leads or transmission lines should be insulated, not only for the normal operating potentials but also sufficiently well to protect against induced voltages caused by lightning hits. A protective gap connected between antenna and ground will discharge heavy induced charges.

HIGH-POWER TRANSMITTERS. High-power transmitters are usually designed so that the equipment is segregated in accordance with a functional grouping, so that the low-power audio-frequency equipment, for example, is well isolated from the high-power r-f amplifiers. Each of these elements has somewhat unique installation problems and will be considered separately.

Audio Equipment. If the transmitter is located some distance from the originating point of the transmitted intelligence and connected to that point by telephone line or radio relay, it is necessary to provide line terminating equipment to match the telephone line properly and to restore the signal, attenuated by the line, to its original amplitude. The necessary pads, line equalizers, line amplifier, together with the percentage-modulation meter and transmitter-frequency monitor, are usually mounted in a standard telephone-type rack in the operating room where they may be conveniently observed by the transmitter operator. It is not ordinarily necessary to provide additional external shielding for this equipment; however, it is necessary that it be adequately grounded through a low-impedance and, preferably, separate ground lead. The incoming audio-frequency circuits and the power supply wiring to the amplifiers should be carefully isolated from each other, preferably by running in separate conduits, or, if in a common trench, by running the audio circuits in conduit. Audio-frequency circuits should use twisted pair enclosed in a tight external shield.

Low-power Intermediate-radio-frequency Amplifiers. This portion of the equipment resembles a low-power transmitter and requires no special comment beyond the points already touched upon under that heading.

High-power Amplifier and Modulator. The installation of these units requires careful planning since they must be installed so that they may be conveniently serviced and so that the required high voltages, cooling air and/or water, and r-f output connections may be conveniently made. Because of the relative simplicity and ease of maintenance, the trend in design of high-power amplifiers has been toward the use of air cooling to dissipate the heat generated in the plate circuits of the r-f amplifiers. For these transmitters provision must be made for an intake for several thousand cubic feet of air per minute, and means must be provided to exhaust this air after it has absorbed the heat developed in the power amplifier tubes. Filters should be provided either in the transmitter or in the air inlet. The exhaust air is frequently piped through ducts out of the transmitter into a mixing chamber where it may either be diverted into the hot air heating system for the building or exhausted outside. Because of the large volume of air required, the inlet duct is frequently installed in the floor below the transmitter. It is possible to use this duct also to carry interconnecting wiring out, but, because of the danger of a rapidly spreading fire in case of an accidental short circuit in this wiring, this practice is not advisable.

Interconnecting wiring between the transmitter units should preferably be carried in separate raceways which may be located in the building floor or arranged as troughs on the side of the transmitter cubicles. All such raceways should be arranged so that they may be conveniently opened for cleaning purposes and to permit changes in wiring as required.

The r-f amplifier should be provided with a low-impedance ground connection preferably in the form of a wide copper strap connected directly to a ground consisting either of buried metal plates or a connection to the antenna ground system, if it is contiguous to the transmitter building.

Rectifier and Power Equipment. It is desirable that the rectifier tubes be located so that they may be observed by the operator at his normal position during the operation of the equipment. Rectifier transformers, however, should be located either in a separate vault or in a transformer yard, enclosed by a grill or fence outside the transmitter house. Fire insurance underwriters' rules should be consulted and closely followed since these rules differ from state to state and usually cover the installation of this type of equipment in some detail.

SUBSTATION. The power line terminal equipment will ordinarily be supplied by the power company having contracted to supply power. This equipment will include any transformers necessary to change the voltage from the power company distribution voltage to the voltage required at the transmitter terminals, together with necessary circuit disconnects, circuit breakers, lightning protectors, and metering equipment. To reduce the possibility of outages resulting from a failure of the power supply, it is advisable, if possible, to provide for power feeds from two separate sources. Automatic equipment is available which will switch from one line to the other in the event of a voltage failure. To secure maximum advantage from such an emergency power supply, the transmitter control circuit should be arranged so that a 1- or 2-sec voltage failure will not necessitate a complete restarting cycle. Either manual or, preferably, automatic reapplication of plate voltage should be provided after a no-voltage drop out, provided, for most rectifier tubes, that such drop out does not exceed 2-sec duration.

BIBLIOGRAPHY

Coleman and Trouant, Recent Developments in Radio Transmitters, *RCA Rev.*, Vol. 3, 316 (January 1939).

Doherty, W. H., A New High Efficiency Power Amplifier for Modulated Waves, *Proc. I.R.E.*, Vol. 24, 1163 (September 1936).

Lee, Reuben, Radio Telegraph Keying Transients, *Proc. I.R.E.*, Vol. 22, 213 (February 1934).

SECTION 8

FREQUENCY MODULATION

FREQUENCY MODULATION

FREQUENCY-MODULATION SYSTEMS

By R. D. Duncan, Jr.

In frequency-modulation systems intelligence is communicated by variation of the frequency or phase of the transmitted wave instead of its strength (as is done in amplitude modulation). Frequency modulation is used for broadcasting, for fixed-to-mobile stations, and for one- or two-way communication in services such as police, public-utility maintenance, and forest patrol. It is being introduced into the truck, bus, taxicab, and railroad fields. So-called "studio-transmitter link" equipment, which provides a one-way radio connection in lieu of telephone lines between the broadcast studio and a remote f-m transmitter, utilizes frequency modulation.

During the war, frequency modulation was much used in long-distance relay service, and it is also used in domestic microwave relay service. It is employed in the allied fields of facsimile and television broadcasting. In the latter, it provides the sound channel and has been employed for relay operation. It has also been experimentally used for power-line carrier current communication.

Frequency modulation is not well adapted to circuits subject to multipath transmission effects since serious distortion results from the simultaneous reception of several signals differing slightly in phase.

1. FUNDAMENTAL RELATIONS

The modulating signal in frequency modulation causes the carrier frequency to be systematically varied above and below the unmodulated value, the extent of the variation being determined by the strength of the signal. The number of times a second the frequency is so varied is determined by the frequency of the signal.

This process of variation of the carrier frequency produces additional frequency components, called sidebands, which lie both above and below the unmodulated carrier frequency. Theoretically, there are an infinite number of such upper and lower sidebands which differ in frequency from the carrier by the value of the modulating frequency or frequencies and integral multiples thereof. However, their amplitudes decrease rapidly as they exceed in frequency value the upper and lower limits of the maximum frequency swing imparted to the carrier so that satisfactory reception can be achieved by a transmitting and receiving pass band somewhat greater than twice this maximum swing. An important characteristic of frequency modulation is that the amplitude of the carrier component as well as of the sideband components is determined by both the amplitude of the modulating signal and by its frequency, or, in the case of a complex wave form, by its component frequencies.

Another distinctive feature of frequency modulation is that, theoretically, the power contained in the carrier and the infinite number of sideband components is a constant value. That is, the transmitted power remains unchanged during the modulation if the circuits are such as to transmit all the sideband frequencies. Practically, since only sidebands within and just greater than the maximum frequency swing are transmitted, there will be slight power variation during modulation, or a small amount of amplitude change or modulation will accompany frequency modulation.

SINGLE-FREQUENCY MODULATION. A consideration of single-frequency modulation will serve to illustrate the essential fundamental characteristics of frequency modulation and the difference between it and phase modulation (see article 8-2 for methods of generating frequency and phase modulation).

When the modulating signal is a single-frequency component, the expression for a f-m voltage may be written in the form

$$e_{\text{f-m}} = E \cos \left(\omega t + \frac{\Delta\omega}{p} \sin pt\right) \tag{1}$$

where E = the maximum amplitude.
ω = 2π times the unmodulated carrier frequency.
p = 2π times the modulating frequency.
$\Delta\omega$ = 2π times the peak frequency swing of the carrier.

The corresponding expression for a p-m voltage is

$$e_{\text{p-m}} = E \cos(\omega t + \Delta\theta \sin pt) \tag{2}$$

where $\Delta\theta$ is the maximum phase or angle variation.

The difference between the two expressions is the ($\sin pt$) term. For frequency modulation the maximum phase or angle excursion is directly proportional to the peak frequency swing, i.e., to the strength of the modulating signal, and inversely proportional to the value of the modulating frequency, whereas for phase modulation, it is proportional only to the strength of the modulating signal and is independent of its frequency.

A physical conception of the difference between frequency and phase modulation may be had from Fig. 1. This shows a voltage vector E which without modulation during an interval of time t has rotated counterclockwise through an angle (ωt). If it is assumed that an observer boards the vector, as it were, so as to rotate with it, and modulation is applied, all that will be noticed is a rocking back and forth of the vector through the angle ($\pm\Delta\theta$), or a rotation first in one direction and then in the other through the angle ($\pm\Delta\theta$), which may greatly exceed 360 deg. If the modulation is purely frequency type, it will be noted that, with a fixed signal strength, the maximum angular excursion will be greater for the low modulating frequencies than for the high ones. If modulation is of the phase type, no difference in the maximum excursion of the vector will be noted for any modulating frequency. If the modulating frequency remains fixed, there is no way of determining whether modulation is frequency type or phase type. If amplitude modulation is also present, a slow periodic shortening and lengthening of the vector would be noticed.

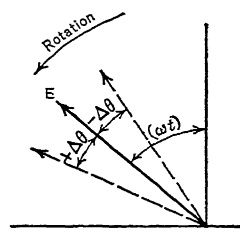


Fig. 1. Voltage Vector with Phase Modulation

In the discussion on the mathematical equivalent of discriminator action (p. 8-09) it is shown that the mathematical equivalent of f-m detection is to take the first differential, with respect to time, of the variable angle through which the vector is rotating. For eqs. (1) and (2) this angle is the argument of the cosine function. Doing this, the two following expressions result.

For frequency modulation

$$\frac{d}{dt}\left(\omega t + \frac{\Delta\omega}{p}\sin pt\right) = \omega + \Delta\omega\cos pt \tag{3}$$

For phase modulation

$$\frac{d}{dt}(\omega t + \Delta\theta \sin pt) = \omega + p\Delta\theta\cos pt \tag{4}$$

The recovered signal is proportional to the periodic term in (3) and (4). For f-m reception, the maximum value of the signal is proportional to the extent of frequency swing and is independent of the value of the modulating frequency. For f-m reception of a phase-modulated wave, the maximum value of the signal increases directly with increasing modulating frequency. By incorporating frequency-distorting circuits at the transmitter or at the receiver, phase modulation may be converted into frequency modulation, or vice versa.

The ratio $\Delta\omega/p$ of Eq. (1) for frequency modulation was originally termed the "deviation ratio" but is now referred to as the "modulation index," the last terminology also applying to the angle $\Delta\theta$ in eq. (2).

The modulation index $\Delta\omega/p$ plays an important part in the theory of noise suppression in f-m systems and involves the circuit characteristics of both the transmitter and receiver. The FCC has established ±75,000 cycles ($\Delta\omega = 2\pi \times 75{,}000$) as 100 per cent modulation for an f-m broadcast transmitter, with a channel band width of 200,000 cycles. It requires that the transmitter be capable of sustaining this maximum peak frequency swing for any aural modulating frequency between 50 and 15,000 cycles, without exceeding certain specified levels of harmonic distortion.

For this maximum peak frequency swing, the modulation index $\Delta\omega/p$ would have a value of 5 for a modulating frequency of 15,000 cycles and a value of 1500 for 50 cycles. There would be five upper and lower sidebands within the overall swing band for the higher modulating frequency and a maximum of 1500 sidebands for the lower frequency.

The standard for 100 per cent modulation for the sound channel of television broadcasting is ±25,000 cycles with the same aural frequency band as f-m broadcasting. It is suggested in the Standards, however, that the f-m transmitter be designed for satisfactory operation at a peak swing of ±40,000 cycles. In the case of point-to-point communication services, such as police, the maximum overall swing band specified by the FCC is three-quarters of the channel band. For the 30–50 megacycle band, one of the several allocated for these services, the channel width is 40,000 cycles, for which the overall swing band would be 30,000 cycles; 100 per cent modulation would then be ±15,000 cycles. Assuming 3000 cycles as the maximum modulating frequency, the modulation index would be 5.

Substituting (P) for the ratio $\Delta\omega/p$ in eq. (1) and for $\Delta\theta$ in eq. (2), the equivalent sideband form of expression may be written as follows:

$$\begin{aligned} e = E[J_0(P)\cos\omega t &- J_1(P)\cos(\omega - p)t + J_1(P)\cos(\omega + p)t \\ &+ J_2(P)\cos(\omega - 2p)t + J_2(P)\cos(\omega + 2p)t \\ &- J_3(P)\cos(\omega - 3p)t + J_3(P)\cos(\omega + 3p)t \\ &+ J_4(P)\cos(\omega - 4p)t + J_4(P)\cos(\omega + 4p)t \\ &- J_5(P)\cos(\omega - 5p)\,t + J_5(P)\cos(\omega + 5p)t \\ &+ \cdots \\ &+ J_n(P)(-1)^n\cos(\omega - np)t + J_n(P)\cos(\omega + np)t] \end{aligned} \tag{5}$$

The coefficients $J_0(P)$, $J_1(P)$, $J_2(P) \cdots J_n(P)$ are Bessel functions of the first kind, of order 1, 2 $\cdots$ n, and argument (P). Values of the argument (P) for frequency modulation, as is shown later on, may vary from approximately 1 radian to the order of 1500 radians. For phase modulation, the maximum value for broadcasting is 5 radians.

For high values of the argument (P), the approximate value of a particular order Bessel function may be computed from the expression (see also Section 1).

$$J_n(P) = \sqrt{\frac{2}{\pi(P)}}\cos\left[(P) - \frac{\pi n}{2} - \frac{\pi}{4}\right] \tag{6}$$

Values of Bessel functions here involved, of zero order, for integral values of the argument from 1 through 9, and for orders 1 through 44, corresponding to integral values of the argument of 1 through 29, are to be found in the book, *Tables of Functions* by Jahnke and Emde. Values corresponding to decimal values of the argument, increasing by increments of 0.2 from 0.2 through 6.0, and in increments of 0.5 from 6.5 through 16.0 for orders zero through 13, are given in British Association for the Advancement of Science, *Reports on the State of Science*, 1915, on The Calculation of Mathematical Tables, pp. 28–33.

For large values of the argument, the reader is referred to Tables of Bessel Functions $J_n(x)$ for Large Arguments by M. S. Corrington and W. Miehle, *Journal of Mathematics and Physics (M.I.T.)*, Vol. 24, 30 (February 1945). Argument values are presented in various incremental groupings, extending from 29 through 300, corresponding to orders zero through 10. Values of the specific function $J_n(1000)$ corresponding to order values of 935 through 1035 increasing by steps of 1 are given in Tables of Bessel Functions $J_n(1000)$ by M. S. Corrington, *Journal of Mathematics and Physics (M.I.T.)*, Vol. 24, 144 (November, 1945).

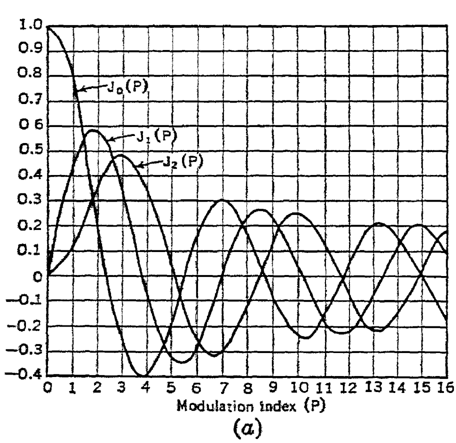
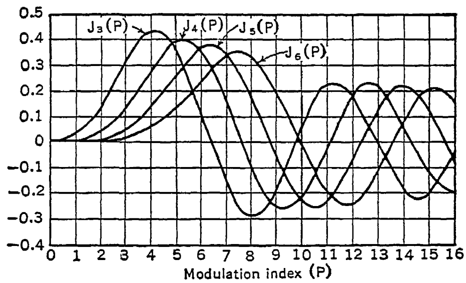


FIG. 2a. Modulation Index (P)

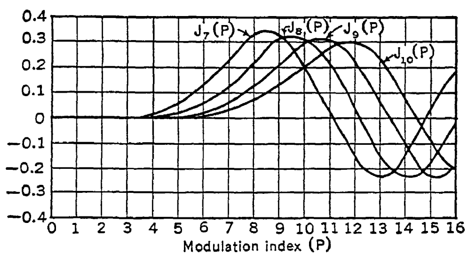
The first ten orders of the Bessel function coefficients of eq. (5) are plotted in Fig. 2. These curves show that, depending upon the value of the argument (P) or modulation indices, $\Delta\omega/p$ or $\Delta\theta$, the different order Bessel functions have both positive and negative values and therefore pass through zero. The amplitude of any sideband component and also of the carrier may be zero, that is, may be entirely missing from the f-m or p-m wave. As is shown later on, this characteristic as it relates to the carrier component provides a basis for measuring the extent of frequency swing or the degree of frequency modulation.

Further illustrative of the relative amplitudes of the carrier and sideband components and the frequency spread, as indicated in eq. (5), values of the first ten order Bessel functions are given in Table 1 for six values of the argument from 0.3 to 5. It is observed that for values of the modulation index of 0.5 and less the amplitudes of the sidebands beyond the first do not exceed approximately 3 per cent of the carrier. For a modulation index



(*b*)

FIG. 2*b*. Modulation Index (*P*)



(*c*)

FIG. 2*c*. Modulation Index (*P*)

of 1.5, the third order sideband is approximately 12 per cent, and the fourth order, 2 per cent, of the carrier. For values of 3 and 5 for the modulation index, all sidebands beyond the sixth and eighth respectively are less than 1 per cent of the carrier amplitude.

Table 1. Values of Bessel Functions

Order n	Argument or Modulation Index (P)					
	(P) = 0.3	(P) = 0.5	(P) = 1.0	(P) = 1.5	(P) = 3.0	(P) = 5.0
0	+0.9770	+0.9380	+0.7650	+0.5050	−0.26010	−0.17760
1	+0.1490	+0.2410	+0.4401	+0.5600	+0.33910	−0.32760
2	+0.0112	+0.0300	+0.1149	+0.2330	+0.48610	+0.04657
3	+0.0006	+0.0025	+0.0195	+0.0615	+0.30910	+0.36480
4			+0.0024	+0.0113	+0.13200	+0.39120
5					+0.04303	+0.26110
6					+0.01139	+0.13100
7					+0.00254	+0.05338
8						+0.01840
9						+0.00250

Equation (5) with Fig. 3 illustrates additional differences between frequency and phase modulation. Figure 3 (*a*) shows a carrier and sideband components for a modulation index equal to 5 radians, and for a modulating frequency of 15,000 cycles. All the components in this and the accompanying figures are plotted as positive regardless of their polarity. For a modulation index of 5 and a frequency of 15,000 cycles, the peak carrier frequency swing is 75,000 cycles. Figures 3(*b*) and (*c*) are for this same peak frequency swing; (*b*) is for an aural frequency of 10,000 cycles and a modulation index of 7.5, and (*c*) an aural frequency of 5000 cycles and modulation index of 15. Figures 3(*d*) and (*e*) are for a fixed value of the modulation index of 5; (*d*) is for an aural frequency of 10,000 cycles and a peak frequency swing of 50,000 cycles, and (*e*) for a 5000-cycle aural frequency and a 25,000-cycle peak frequency swing. Figure 3(*a*) is representative of both frequency and phase modulation. Figures 3(*b*) and (*c*) are representative of frequency modulation and Fig. 3(*d*) and (*e*) of phase modulation.

For frequency modulation a decrease in the modulating frequency with a fixed peak frequency swing increases the peak angle swing, alters the respective values of the carrier and sideband components, increases the number of important sidebands within the frequency swing band, and bunches the sidebands lying just outside of the swing band closer

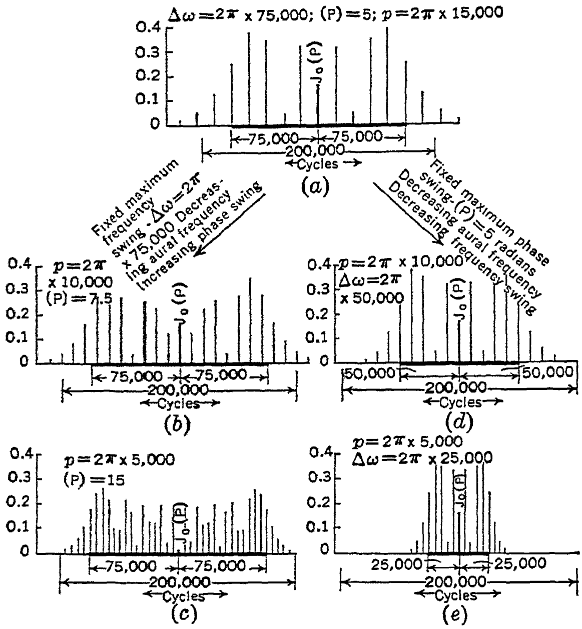


FIG. 3. Sideband Distribution for FM and PM

to the peak swing limits. For the three aural frequencies considered, the first three sidebands lying just beyond the swing limits, the outer one of which is less than 5 per cent in value of the unmodulated carrier, occupy frequency spaces of respectively 45,000, 25,000 and 15,000 cycles.

Going to an extremely low modulating frequency, for example, 75 cycles, there would be a maximum of 1000 upper and lower sidebands within the swing band with some appreciable out-of-band components bunched relatively close to the swing limits. Mathematical analysis of this particular case has shown that the fifteenth out-of-band component which would be displaced by 1125 cycles beyond the swing limits is approximately 1 per cent in amplitude of that of the unmodulated carrier, with the further outlying sidebands rapidly approaching zero.

It is to be concluded for frequency modulation that (*a*) for a given energy content the higher modulating frequencies require a somewhat greater band-pass range from a circuit design standpoint and (*b*) with decreasing modulating frequency the required band-pass range approaches the overall frequency swing band in width. This last explains why only the actual swing band need be considered when the frequency is varied at a very slow rate, for example, by manual change of the capacitance of an oscillator, even though variation is over a relatively large range. It also explains why steady-state analysis, under the assumption of a sinusoidal frequency variation at a low rate, may be employed in the study of f-m circuits yielding results which, when properly interpreted, are sufficiently accurate for design purposes (see article 2-8).

Furthermore the fact that in the usual speech or music programs the energy of these high-frequency tones is quite low keeps the modulation low and hence the magnitude of harmonics of higher order.

The practical result of these factors is that f-m systems work very well with the pass band only slightly greater than the swing band.

In the case of phase modulation, decreasing the modulating frequency, with a fixed phase swing, decreases the overall frequency swing, that is, bunches all sidebands closer together, but alters neither the number of sidebands, their respective amplitudes, nor the amplitude of the carrier.

If eq. (5) represents an f-m voltage wave, the rms value is given by

$$E_{\mathrm{rms}} = \frac{E}{\sqrt{2}}\left[J_0^2(P) + 2J_1^2(P) + 2J_2^2(P) + 2J_3^2(P) + 2J_4^2(P) + 2J_5^2(P) + \cdots\right]^{1/2} \tag{7}$$

The theory of Bessel functions shows that the quantity in the brackets approaches unity as the number of terms becomes large. That is, the rms voltage of an f-m wave or the power transmitted remains substantially at a constant value regardless of the extent of modulation if a sufficient number of sidebands are present.

TWO-FREQUENCY MODULATION. Corresponding to eq. (1), the expression for two-frequency modulation may be written in the form

$$e_{\text{f-m}} = E \cos\left(\omega t + \frac{\Delta\omega}{p}\sin pt + \frac{\Delta\omega}{q}\sin qt\right) \tag{8}$$

where E, ω, p, and $\Delta\omega$ have the same meanings as before and $q = 2\pi$ times the second modulating frequency. The equivalent sideband form of this expression is given by eq. (9), in which, as before, $(P) = \Delta\omega/p$ is the modulation index of the $p/2\pi$ frequency, and (Q), substituted for $\Delta\omega/q$, is the modulation index of the $q/2\pi$ frequency. For simplification, this expansion is limited to the third-order Bessel functions of the two arguments.

The result of two-frequency modulation is the production of major sidebands involving each of the two frequencies and, in addition, other sidebands involving the sums and differences of the two frequencies and integral multiples of the same in different combinations. The amplitudes of all sidebands are products of pairs of Bessel functions. Since, for finite values of the argument, the values of all Bessel functions are less than unity, the product of any pair will be less than the individual value of either of the pair. The change from single- to two-frequency modulation thus greatly increases the number of sideband components and the frequency spread or space occupied. However, the amplitudes of all sidebands have been decreased in value so that, while the frequency spread has increased, the outer sidebands have very low amplitudes.

A similar situation holds when more modulating components are added. That is, in going from a single sinusoidal modulating signal to one of complex wave form the number of sideband components is greatly increased, but their amplitudes automatically become readjusted, or the total modulation is in effect divided between the components, so that the overall frequency spread for all practical purposes undergoes little or no change.

$$\begin{aligned}
e = E[J_0(P)J_0(Q)\cos\omega t &- J_1(P)J_0(Q)\{\cos(\omega - p)t + \cos(\omega + p)t\} \\
&+ J_2(P)J_0(Q)\{\cos(\omega - 2p)t + \cos(\omega + 2p)t\} \\
&- J_3(P)J_0(Q)\{\cos(\omega - 3p)t + \cos(\omega + 3p)t\} \\
&\cdot \\
&\cdot \\
&\cdot \\
&- J_0(P)J_1(Q)\{\cos(\omega - q)t + \cos(\omega + q)t\} \\
&+ J_0(P)J_2(Q)\{\cos(\omega - 2q)t + \cos(\omega + 2q)t\} \\
&- J_0(P)J_3(Q)\{\cos(\omega - 3q)t + \cos(\omega + 3q)t\} \\
&\cdot \\
&\cdot \\
&\cdot \\
&- J_1(P)J_1(Q)\{\cos[\omega - (p - q)]t + \cos[\omega + (p - q)]t \\
&\qquad - \cos[\omega - (p + q)]t + \cos[\omega + (p + q)]t\} \\
&- J_1(P)J_2(Q)\{\cos[\omega - (p - 2q)]t - \cos[\omega + (p - 2q)]t \\
&\qquad + \cos[\omega - (p + 2q)]t - \cos[\omega + (p + 2q)]t\} \\
&- J_1(P)J_3(Q)\{\cos[\omega - (p - 3q)]t + \cos[\omega + (p - 3q)]t \\
&\qquad - \cos[\omega - (p + 3q)]t - \cos[\omega + (p + 3q)]t\} \\
&\cdot \\
&\cdot \\
&\cdot
\end{aligned}$$

$$
\begin{aligned}
&+ J_2(P)J_1(Q)\{\cos[\omega - (2p - q)]t - \cos[\omega + (2p - q)]t \\
&\qquad - \cos[\omega - (2p + q)]t - \cos[\omega + (2p + q)]t\} \\
&+ J_2(P)J_2(Q)\{\cos[\omega - (2p - 2q)]t + \cos[\omega + (2p - 2q)]t \\
&\qquad + \cos[\omega - (2p + 2q)]t + \cos[\omega + (2p + 2q)]t\} \\
&+ J_2(P)J_3(Q)\{\cos[\omega - (2p - 3q)]t + \cos[\omega + (2p - 3q)]t \\
&\qquad - \cos[\omega - (2p + 3q)]t + \cos[\omega + (2p + 3q)]t\} \\
&\cdot \\
&\cdot \\
&\cdot \\
&- J_3(P)J_1(Q)\{\cos[\omega - (3p - q)]t + \cos[\omega + (3p - q)]t \\
&\qquad - \cos[\omega - (3p + q)]t + \cos[\omega + (3p + q)]t\} \\
&- J_3(P)J_2(Q)\{\cos[\omega - (3p - 2q)]t - \cos[\omega + (3p - 2q)]t \\
&\qquad + \cos[\omega - (3p + 2q)]t - \cos[\omega + (3p + 2q)]t\} \\
&- J_3(P)J_3(Q)\{\cos[\omega - (3p - 3q)]t + \cos[\omega + (3p - 3q)]t \\
&\qquad - \cos[\omega - (3p + 3q)]t - \cos[\omega + (3p + 3q)]t\} \\
&\cdot \\
&\cdot \\
&\cdot \\
&+ \cdots]
\end{aligned} \tag{9}
$$

To avoid overmodulation on the peaks, the average modulation with a complex wave form must be reduced. The necessity for such a reduction is well known in broadcasting.

MEASUREMENT OF FREQUENCY SWING. In the development, testing, and practical operation of f-m equipment, it is necessary to be able to measure the extent of frequency or phase swing or modulation. The fact that the zero-order Bessel function $J_0(P)$ passes through zero with increasing values of the modulation index as shown in Fig. 2 is made the basis for one useful method of measuring the peak frequency or phase swing. In Table 2 are given the first five values of the argument (P) or modulation index for which the zero-order Bessel function $J_0(P)$ becomes zero. Other values may be obtained from the references previously given.

Table 2. First Five Values of Argument (P) for Which $J_0(P) = 0$

2.4048
5.5201
8.6537
11.7915
14.9309

In the operation of this method, a receiver of the heterodyne type is tuned to the unmodulated carrier so that a beat tone of several hundred cycles is obtained. Single-frequency modulation is then applied to the transmitter, the frequency being of a much higher order than that of the beat tone. The strength of the modulating signal is increased until the amplitude of the beat tone drops to zero, which for the first null corresponds to the first passage through zero of the zero-order Bessel function $J_0(P)$. From Table 2, the value of the modulation index for this null point is 2.4048. The value of the modulating frequency $(p/2\pi)$ being known, the peak frequency swing $(\Delta\omega/2\pi)$ is obtained from the ratio $\Delta\omega/p = 2.4048$. Further increase in strength of the modulating signal will develop additional null points corresponding to which values of the modulation index are given in Table 2.

Another method of indirectly measuring the frequency swing provides a visual observation on the screen of a cathode-ray oscilloscope, of the carrier and sideband components in their respective amplitudes and positions within a frequency band somewhat greater than twice the swing band. In this method, the f-m signal wave is heterodyned to an intermediate frequency of 2 Mc, amplified, passed through an a-m detector, and supplied to the vertical deflection plates of a cathode-ray oscilloscope. The heterodyning oscillator is itself frequency-modulated by means of a reactance tube, with a 25-cycle linear-sweep signal to peak values of plus and minus 100,000 cycles. The linear-sweep voltage is also impressed on the horizontal deflection plates of the cathode-ray oscilloscope.

In action, the f-m signal wave, which consists of its carrier and sideband components, is in effect "scanned" frequency-wise from 100,000 cycles below to 100,000 cycles above the carrier, at a rate of 25 times a second. When, during this scanning process, a sideband component or the carrier is encountered, a voltage pulse is impressed on the vertical deflection plates of the oscilloscope. As the linear-sweep voltage is also impressed on the horizontal plates, the particular component will appear in its proper location within the overall swing band.

THE MATHEMATICAL EQUIVALENT OF DISCRIMINATOR ACTION. In theoretical investigations of f-m systems, as with other communication systems, the end product usually sought is a determination of the form of the received signal in comparison with that of the signal originating at the transmitter.

An alternating voltage or current is represented by a vector of length proportional to the maximum or rms value, revolving at an angular velocity equal to 2π times the frequency, or, in customary nomenclature, $\omega = 2\pi f$. In a given time t, the vector will have rotated through an angle of $(\omega t = 2\pi\, ft)$ radians. From the elementary theory of the mechanics of planer rotating bodies, it is known that the angular velocity of rotation of any point in the body about the axis is equal to the first differential of the instantaneous angle of rotation with respect to the time. In the present case, this would be $d(\omega t)/dt = \omega = 2\pi f$; that is, the frequency is equal to the first differential of the instantaneous angle with respect to time.

In conventional a-c theory, a vector revolves at a constant velocity as the frequency has a constant value. In frequency modulation, the angular velocity varies with time, in accordance with the modulation, about a mean or unmodulated value. The frequency obtained by differentiation is therefore the "instantaneous frequency," which is what is sought. The expression for the instantaneous frequency may be a simple periodic function of the modulating frequency or it may be a complex function requiring algebraic and trigonometric operation or a Fourier analysis to break it down into the fundamental and harmonic components.

Expressions are given above for a single- and a double-frequency-modulated f-m voltage wave and single-frequency-modulated p-m voltage wave. These are as follows:

$$e_{\text{f-m}} = E \cos\left(\omega t + \frac{\Delta\omega}{p} \sin pt\right)$$

$$e_{\text{f-m}} = E \cos\left(\omega t + \frac{\Delta\omega}{p} \sin pt + \frac{\Delta\omega}{q} \sin qt\right)$$

$$e_{\text{p-m}} = E \cos\left(\omega t + \Delta\theta \sin pt\right)$$

The first differential with respect to the time of the arguments of the three cosine functions are, respectively, $(\omega + \Delta\omega \cos pt)$ and $(\omega + \Delta\omega \cos pt + \Delta\omega \cos qt)$ and $(\omega + p\,\Delta\theta \cos pt)$. With linear detection the demodulated output is directly proportional to these expressions. With square-law detection, the demodulated output is proportional to the squares of the expressions and would contain second-harmonic components in addition to the fixed and fundamental frequency components. With a balanced discriminator and differentially connected detectors, both the fixed and second-harmonic components cancel out, leaving the fundamental frequency components of double amplitudes. Linear detection is employed almost exclusively in commercial practice.

It is important to note that this mathematical method of recovering an f-m signal implicitly assumes (*a*) that there is no amplitude modulation and (*b*) a linear input-output characteristic of the hypothetical discriminator. It gives only the relative amplitudes of the frequency components but tells nothing of their absolute values. If amplitude modulation is present, the maximum voltage amplitude E in the preceding expressions is also a function of the time and must be taken into due account. Under most practical conditions, it may be assumed that the effects of amplitude modulation are removed, by some means such as limiting, leaving only the variable angle to be investigated.

FREQUENCY-MODULATION TRANSMITTERS

By J. E. Young

Transmitters for two major types of services are discussed in this section. These are frequency-modulation broadcasting, which includes the sound transmitters used for television broadcasting, and emergency communication.

2. FREQUENCY-MODULATION BROADCASTING

Frequency-modulation broadcasting has been assigned the frequency range from 88 to 108 mc. This band has been divided into 100 contiguous channels with carrier frequencies starting at 88.1 mc and ending at 107.9 mc. Transmission in each channel is permitted with a maximum frequency swing of 75 kc, which is designated 100 per cent modulation. To insure that f-m broadcasting will be a high-fidelity service, the FCC has set up the following standards for the overall transmitting system from microphone terminals to antenna:

1. Distortion: less than 3.5 per cent, 50 to 100 cycles, 2.5 per cent, 100 to 7500 cycles at 25 per cent, 50 per cent, and 100 per cent modulation; and 3.0 per cent, 7500 to 15,000 cycles at 100 per cent modulation.

2. Noise level: at least 60 db below 100 per cent modulation in the band 50 to 15,000 cycles.

3. Amplitude noise: at least 50 db below 100 per cent amplitude modulation in the band 50 to 15,000 cycles.

4. Frequency characteristic: the transmitting system shall be capable of transmitting a band of frequencies from 50 to 15,000 cycles. Pre-emphasis shall be employed in accordance with the impedance-frequency characteristic of a series inductance-resistance network having a time constant of 75 microseconds. The deviation of the system response from the standard pre-emphasis curve shall be between two limits. The upper of these limits shall be uniform (no deviation) from 50 to 15,000 cycles. The lower limit shall be uniform from 100 to 7500 cycles, and 3 db below the upper limit; from 100 to 50 cycles, the lower limit shall fall from the 3-db limit at a uniform rate of 1 db per octave (4 db at 50 cycles); from 7500 to 15,000 cycles the lower limit shall fall from the 3-db limit at a uniform rate of 2 db per octave (5 db at 15,000 cycles).

Included in the system for which performance is thus specified are the microphone preamplifier, mixers, program amplifier, studio-to-transmitter link, which may be wire or radio, transmitter line terminating amplifier, and transmitter. The Radio Manufacturers Association, working through industry committees, has specified the performance of the components of the system so that it would be possible to combine elements of different manufacture without exceeding the distortion or noise limits specified. For the transmitter, the following standards have been agreed to in the industry:

1. Distortion: audio distortion, including all harmonics up to 30 kc, shall not exceed 1.5 per cent rms from 50 to 15,000 cycles, and shall not exceed 1 per cent rms between 100 and 7500 cycles. Measurements shall be made at 25 per cent, 50 per cent, and 100 per cent modulation, for audio frequencies of 50, 100, 400, 1000, and 5000 cycles, also at 100 per cent modulation for audio frequencies of 7500, 10,000, and 15,000 cycles.

2. Noise level: at least 65 db below 100 per cent modulation in the band 50 to 15,000 cycles.

3. Amplitude noise: at least 50 db below 100 per cent amplitude modulation in the band 50 to 15,000 cycles.

4. Frequency characteristic: shall not deviate more than 1 db from a straight line, or, if pre-emphasis is used, from a 75-microsecond curve from 50 to 15,000 cycles.

F-M TRANSMITTER—SCOPE. Like the a-m transmitter covered in pp. 7-128, 7-137, the f-m transmitter is considered to consist of all audio equipment operating above standard telephone-practice level and all r-f equipment from the source of the r-f oscillation to the transmission-line terminals. Unlike a-m transmitters, f-m practice is to modulate at low power levels, multiplying the frequency and power, usually many times, before reaching the transmitter output.

FREQUENCY CONTROL AND MODULATION. Frequency control and modulation are tied together since modulation is accomplished by actually swinging the frequency back and forth in accordance with the modulating signal. To accomplish this, and still keep the center, or average, frequency within tolerance (±2000 cycles for f-m broadcasting) is one of the chief problems of f-m transmitter design.

Methods of frequency modulation may be divided into two basic systems. In one of these, phase modulation, modulation is effected at some point in the circuit following the oscillator, which is crystal-controlled. Modulation is accomplished by changing the phase of the crystal-controlled signal at a rate corresponding to the desired modulation. It is characteristic of p-m systems that the frequency shift is proportional to the modulating frequency as well as to its amplitude. To convert to true frequency modulation it is necessary to compensate the frequency characteristic of the modulator, therefore, so that the amplitude of the modulating signal is inversely proportional to its frequency. One system of phase modulation is shown in the block diagram in Fig. 1. A crystal oscillator having a frequency of approximately 200 kc is used to drive a balanced modulator in which the modulating signal is the audio frequency to be transmitted, and whose a-f characteristic has been corrected to convert from phase to frequency modulation as described above. The output of the balanced modulator consists of two f-m currents whose deviations are instantaneously in opposite directions. These two signals are multiplied in frequency 81 times through two separate channels of multipliers. One of the resultant signals is heterodyned to another frequency about 2 mc removed. The resultant frequency and the output of the second multiplier channel are then recombined and their difference recovered. It will be noted that the difference frequency is independent of the frequency of the original oscillator and depends only on the frequency of the 2-mc heterodyne oscillator. Additional multiplication of 48 times is then necessary to reach the final operating frequency.

The phase modulation, at the point of its introduction, is usually restricted to less than 30° in order not to exceed permissible transmitter distortion. In the method described

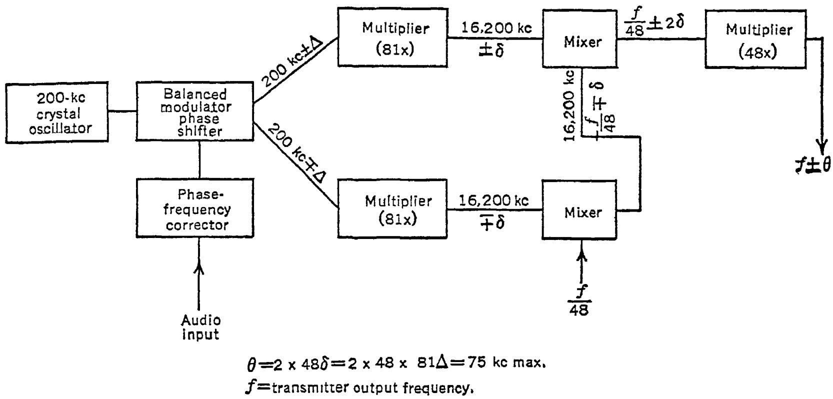


FIG. 1. Frequency Modulation by the Phase-modulation Method

above, since two outputs are obtained from the balanced modulator having phase modulation in the opposing sense, a total phase shift of 60° is possible. The multiplication necessary to convert this phase modulation to frequency modulation, having a frequency swing of ±75 kc, was, in this particular example, 3888 times. The use of dual p-m channels not only permits doubling the frequency swing but also makes it possible to cancel out the effect of the 200-kc crystal in determining the transmitter frequency stability. In the arrangement shown, the carrier frequency stability is a function only of the stability of the 2-mc oscillator. Some improvement in f-m noise level is also obtained at the same time.

Other methods of phase modulation have been developed which are quasi-mechanical in nature. One of these, for which a special vacuum tube has been developed known as the "Phasitron" tube, is used in some commercial f-m transmitters. The block diagram of the circuit is shown in Fig. 2. A crystal oscillator operating at approximately 230 kc

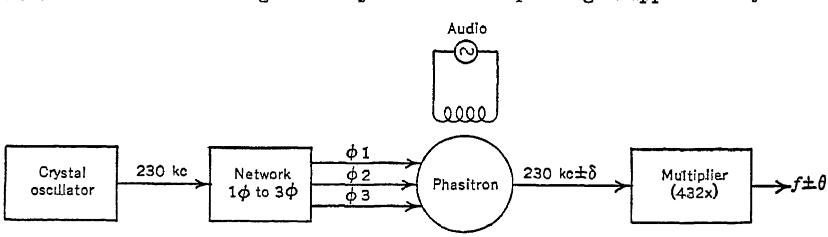


FIG. 2. Frequency Modulation Using the "Phasitron" Tube

(the exact frequency depends on the assigned station frequency) is coupled to a network which, by means of tuned circuits, produces three outputs, all of the same frequency and separated in phase by 120°. These outputs are applied to alternate wires of the deflector grid of the Phasitron tube in such a manner as to produce a rotating field. A cathode and two anodes, which are at a positive d-c potential with respect to the cathode, are arranged so that electrons are drawn from the cathode and focused into the form of a tapered thin-edged disk. This disk, with the cathode for its axis, lies between a neutral plane and the deflector grid structure and extends out to anode 1. The three-phase potential applied to the deflector grid structure deflects the electron beam so that the outer edge of the electron "disk," if it could be seen, would appear to have a sinusoidally serrated edge which rotates about the cathode as a center, at a speed determined by the three-phase voltage applied to the grid. Anode 1, located cylindrically around the cathode, outside the periphery of the deflector grids, has 24 evenly spaced holes arranged around its circumference; 12 of these are above the plane of the electron disk and 12 below. The rotating serrated edge of the electron disk impinges on this series of holes. As the serrations move around the periphery of the electron disk, the electrons alternately pass through the holes to anode 2 and as the serrations move on one-half cycle are completely blocked from anode 2. Thus the current flowing to this anode varies sinusoidally at the crystal frequency. It follows, therefore, that any variation in the angular velocity of rotation of the electron disk will result in a phase variation of the output current from anode 2.

A coil is placed around the outside of the tube so that the magnetic field resulting from the current flowing in the coil is perpendicular to the plane of the electron disk. The electrons traveling radially out of the cathode toward the anodes through this field have a force exerted on them in a direction perpendicular to their path and perpendicular to the direction of the magnetic field. Thus, an angular displacement is introduced in the rotation of the electron disk causing phase variation in the output of anode 2 as described above. Consequently, if an a-f signal current of the proper amplitude flows through this coil the output of the tube will be phase-modulated in accordance with this audio frequency. When there is no audio signal input, the serrations about the periphery of the electron disk will rotate at a constant amplitude, and the output frequency of anode 2 will be the same as the frequency of the crystal.

Phase excursions as great as 720° are possible, in this system, but if low distortion is to be achieved the phase shift must be considerably restricted.

DIRECT FREQUENCY MODULATION. Direct modulation of the transmitter master oscillator may be accomplished by several different methods. They all function by changing the reactance of the frequency-controlling part of the oscillator circuit

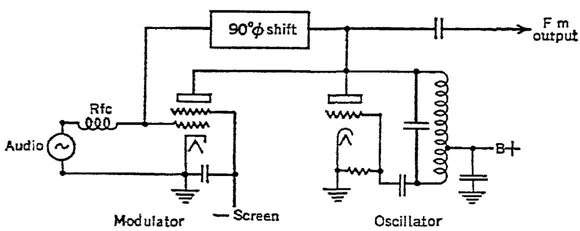


FIG. 3. Frequency Modulation by Reactance Tubes

in accordance with the modulating frequency. One of the commonest of such systems is the use of a tube or tubes having their plates connected to the oscillator circuit and their grids excited by an r-f voltage derived from the oscillator tank and shifted in phase 90° with respect to the oscillator tank voltage. Plate current drawn by these tubes will then be 90° out of phase with the oscillator output and thus has the characteristic of a positive or negative reactance. The amplitude of this plate current is then varied by means of the audio signal, which is also impressed on the grid, and, consequently, the oscillator frequency shifts in accordance with the a-f signal. The basic circuit is shown in Fig. 3.

Another method of achieving frequency shift in the master oscillator is to connect the grid circuit of the modulator in parallel with the oscillator tank. See Fig. 4. In this cir-

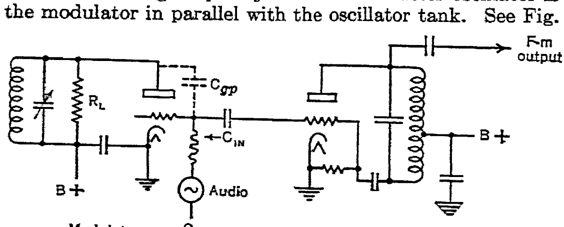


FIG. 4. Frequency Modulation by Input Capacitance Variation

cuit, frequency modulation results from the fact that the input capacitance of the modulator tube is a function of its plate circuit impedance and its transconductance, so that

$$C_{in} = C_{gp}(1 + g_m R_L)$$

It follows that, if the transconductance is varied by the audio input signal, frequency modulation in accordance with the audio input is obtained.

All the systems of direct frequency modulation have the common characteristic that the transmitter frequency control is effected with reference to a separate, highly stable oscillator. These control systems may be divided into two categories, those in which the restoring force is proportional to the deviation, and those in which full restoring force is developed regardless of the amount of deviation. In general, the latter will provide the most accurate frequency control since the transmitter output frequency will be an exact multiple of the reference oscillator frequency if the system is functioning properly. The control voltage developed by the error signal in either system may be used to correct the master oscillator frequency, either by changing the bias of the tube used to effect modulation or through an electromechanical device driving a correcting capacitor in the tank circuit of the oscillator.

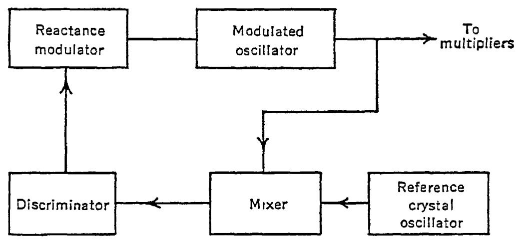


Fig. 5. The Crosby Method of Frequency Control

One of the earliest of the first type of systems is shown in Fig. 5. A portion of the output of the f-m oscillator, which is modulated by reactance tubes, is fed into a mixer, together with the signal from the reference crystal oscillator. The beat between these two signals is fed into a discriminator tuned so that it is working on the center of its characteristic when the frequency of the modulated oscillator is the exact submultiple of the transmitter frequency. As the modulated oscillator drifts away from the correct frequency, the beat signal fed into the discriminator changes frequency accordingly and a d-c voltage is developed by the discriminator. This voltage is applied to the grids of the reactance tubes in the proper sense to correct the frequency of the oscillator. This method of frequency correction may be thought of as akin to negative feedback, since the degree of correction is a function of the gain through the mixer and discriminator, and the sensitivity of the reactance tubes.

One of the several variations of the second type of system is shown in Fig. 6. The master oscillator is frequency-modulated by a reactance tube or other electronic means. A sample of its output is divided in frequency by a locked-in oscillator, or multivibrator, to a frequency low enough so that the carrier frequency will not vanish for any percentage of modulation of any audio frequency in the normal pass band. For example, in one transmitter the modulated oscillator has a center frequency of 4 Mc, requiring a 3-kc swing to achieve a 75-kc frequency swing at the transmitter output frequency. The center frequency is divided by 256 so that the frequency swing reaches a maximum of 12 cycles per second. If an audio modulating frequency of 30 cycles per second is used, the phase shift is then 0.4 radian or 24°. This modulation index reduces the carrier to 0.96 times its unmodulated value; consequently, a carrier frequency component is always present to effect synchronization. The output of the reference crystal oscillator is also divided in frequency, in order to use a crystal which can be easily manufactured, and these two outputs are fed into a balanced phase detector. If the two inputs to the phase detector are exactly 90° out of phase its output will be zero. Any shift in phase difference away from the 90° point will produce an output voltage with a positive or negative sign depending on whether the phase shift is greater or less than 90°. This voltage is then applied to the modulator tube grid to correct the frequency of the master oscillator.

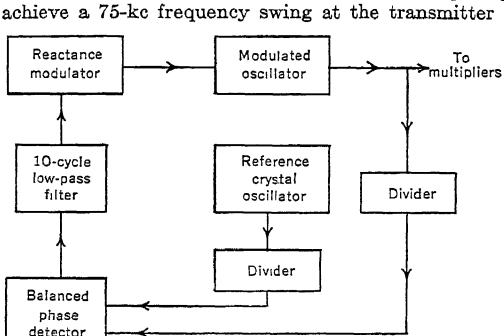


Fig. 6. Phase Detector Method of Frequency Control

A system of frequency correction utilizing an electromechanical circuit is shown in block diagram in Fig. 7. In this circuit the oscillator is modulated by reactance tubes in the conventional manner. A sample of the output of the modulated oscillator is divided by 240 and then fed into the grids of a pair of balanced modulators. The output of the reference crystal oscillator is divided by 5 and then split through two phase-shifting networks, arranged to provide a 90° phase difference between the two branches. One of these branches is fed into the grid circuit of one of the balanced modulators and the other to the grid circuit of the second balanced modulator. If the divided down signal from the f-m oscillator is in phase with the divided signal from the reference oscillator there will be no a-c output from the balanced modulators. However, if there is any frequency difference, an a-c voltage will be developed in the plate circuits of each of the balanced modulators and the two voltages thus obtained will differ in phase by 90°. These

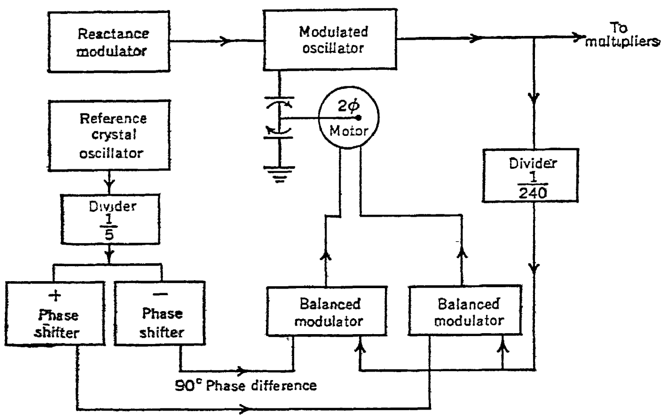


FIG. 7. Electromechanical Frequency Control

two outputs are connected to the two windings of a two-phase motor, and the shaft of this motor is arranged to drive a small capacitor connected in the tank circuit of the f-m oscillator. The a-c outputs of the balanced modulators, resulting from a frequency deviation in the modulated oscillator, produce a rotating field in this motor which tends to rotate the variable capacitor in the proper direction to correct for the frequency error. As soon as the frequency of the modulated oscillator has been restored to the point where its submultiple is in exact synchronism with the submultiple of the crystal oscillator, the a-c output of the balanced modulators drops to zero and the tuning capacitor comes to rest.

Another frequency-control circuit operating on a different principle is shown in the block diagram, Fig. 8. In this circuit the outputs of the master oscillator and the crystal oscillator are combined in a pair of mixers, in one of which the input from the crystal oscillator differs in phase by 90° from the input to the other. The output of one of the

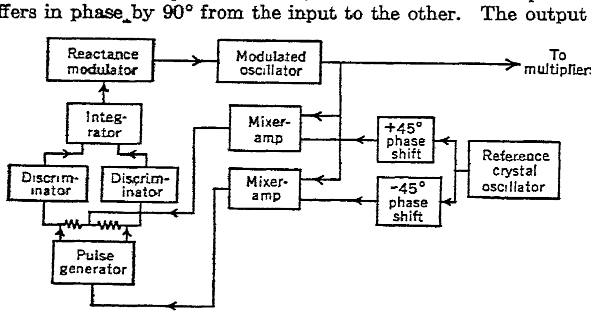


FIG. 8. Pulse Counter Frequency Control

mixers is fed through a pulse generator and thence through a pair of discriminators. The discriminators are biased diodes. The bias on these diodes is set just above the peak value of the output of the second mixer. The result is that, when the pulses add to the sine-wave output of the second mixer, the bias is overcome and the pulse is passed through the diode. When the pulse subtracts from the sine wave the bias prevents the diode from conducting and the pulse is not passed. This arrangement serves to separate the pulses into two circuits; one circuit is energized by one pulse for each cycle of beat between the master oscillator and the crystal oscillator when the signal frequency is high, and the other circuit is energized by one pulse for each cycle of beat when the signal frequency is low. The outputs of the two discriminators are fed to an integrator, which is simply a large capacitance and thence to the grid circuit of a cathode follower. The output of this tube is connected to the tube which effects frequency modulation, and thereby controls the frequency of the master oscillator. This circuit will tend to hold the center frequency so that the frequency swings higher and lower than the correct frequency by the same total number of cycles. The correction is continuous, and the speed at which a frequency shift is reflected in a correcting voltage is a function of the time constant of the integrating capacity.

3. FREQUENCY MODULATION FOR EMERGENCY TRANSMITTERS

Phase modulators, corrected to obtain frequency modulation, are used almost exclusively for this class of service. The necessary degree of phase modulation can be achieved with relatively few stages of multiplication of the f-m signal. This results from the combination of two requirements that greatly restrict the necessary p-m angle. The first of these is that a comparatively narrow frequency swing is used, varying from ± 12.5 kc in the 25 to 30 Mc band to ± 22.5 kc in the 152 to 162 Mc band. The second is that the lowest unattenuated modulating frequency need be no lower than 500 cycles per second. Thus, if a phase swing of 1 radian can be obtained in the modulator, a multiplication of only 45 is necessary to achieve a frequency swing of 22.5 kc. Since considerable distortion is tolerable before any loss of intelligibility results, phase modulators having much more inherent distortion may be used for this class of service than for f-m broadcasting. Overall distortion as high as 10 to 15 per cent has been found to have no effect on the intelligibility of the signal. Since most transmitters of this type are portable, the most important characteristics are small size and weight, and the minimum number of tubes.

4. TRANSMITTER CIRCUITS

Because of the many times the f-m signal must be multiplied, in f-m transmitters, to obtain the necessary frequency swing, the low power

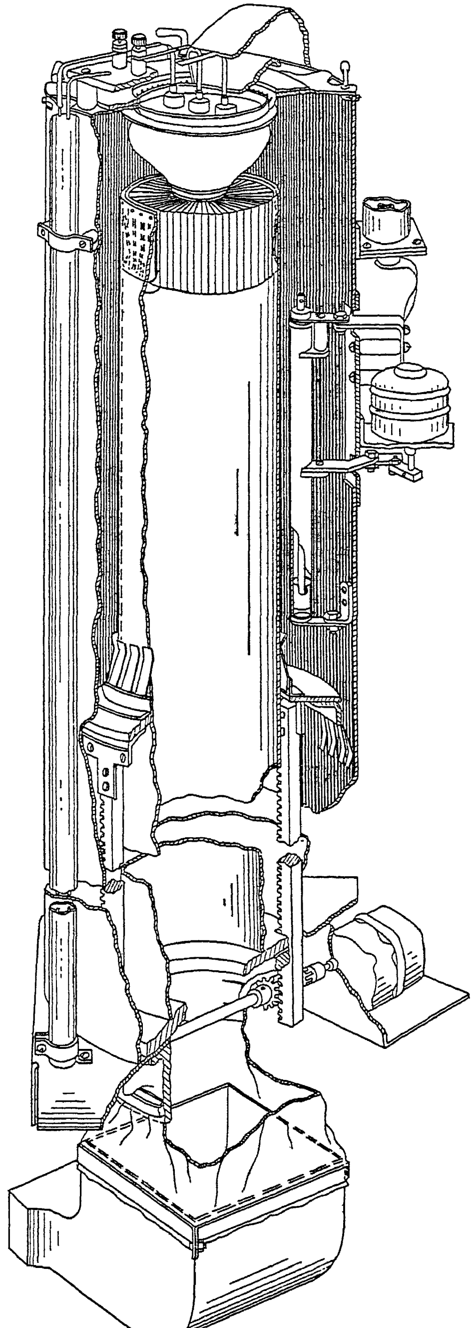


FIG. 9. Coaxial Tank Circuit

stages operate through the frequency range covered by the section on a-m transmitters. Additional care is required in f-m transmitter design to avoid high-Q circuits, since the band widths required are greater than in other transmitters. For the same reason, it is desirable to use circuits that have symmetrical phase-shift characteristics about the center frequency.

Output amplifiers for transmitters of power above 250 watts usually employ tubes having internal capacitances great enough to necessitate the use of transmission-line-type tanks. To avoid stray fields which might affect the operation of the exciter stages, and to simplify the problem of keeping the transmitter enclosure and outer conductor of transmission lines at ground potential, these circuits are preferably made in the form of concentric lines. One such typical circuit is shown in Fig. 9. In this illustration a triode is used. To avoid interaction between plate and grid circuits, the grid is grounded for r-f voltages, and excitation voltage is applied between ground and the filament of the tube. A three-quarter wave tuned transmission line is used and is in turn coupled to the output of the driver by a small, single-turn loop. The plate circuit is a coaxial line, in which the anode of the tube forms a continuation of the inner conductor. The line is tuned by moving a by-pass capacitor provided with fingers which make contact with the inner and outer conductor, along the line. Output coupling is obtained by positioning a loop in the space between the inner and outer conductors. Control of the tightness of the coupling is effected by changing the angle of the loop. Maximum coupling is obtained when the loop lies along a radius of the outer conductor.

BIBLIOGRAPHY

Carson, John R., Notes on the Theory of Modulation, *Proc. I.R.E.*, Vol. 10, 57 (February 1922).

Armstrong, Edwin H., A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation, *Proc. I.R.E.*, Vol. 24, 689 (May 1936).

Crosby, Murray G., Band Width and Readability in Frequency Modulation, *RCA Rev.*, Vol. 5, 363 (January 1941).

Jaffe, David Lawrence, A Theoretical and Experimental Investigation of Tuned Circuit Distortion in Frequency Modulation Systems, *Proc. I.R.E.*, Vol. 33, 318 (May 1945).

FREQUENCY-MODULATION RECEIVERS

By Leslie F. Curtis

5. COMPARISON WITH AMPLITUDE-MODULATION RECEIVERS

Conventional f-m receivers use the superheterodyne principle and differ from a-m superheterodynes mainly in the i-f amplifier and in the second detector.

The intermediate frequency is chosen to give adequate image reduction depending on the service. An intermediate frequency of 10.7 Mc is suitable, and is specified as an RMA standard, in receivers for f-m broadcasting in the assigned band from 88 to 108 Mc since the image response then falls outside the bands having assigned services which are liable to interfere. A higher intermediate frequency of the order of 21.7 Mc is required for the sound channels of television receivers to avoid interference. The higher frequency is favored in receivers incorporating both f-m broadcast and television facilities since the same components may then be used for both.

The i-f band width should be sufficient to pass the side frequencies at the maximum system frequency deviation without excessive attenuation of the power in any of them to preserve the fidelity of the modulation. A band width of at least 150 kc for a reduction of not over 50 per cent on the overall response curve at maximum deviation is required in receivers for f-m broadcasts. (See Section 7, article 12.) This applies even when an amplitude limiter follows the i-f amplifier, since, although a limiter removes amplitude modulation, it cannot remove the phase distortion introduced in the tuned circuits. A flat or slightly rounded i-f response curve is preferred since double-peaked curves increase the phase distortion within the receiver. The uniformity of response over the required band should be as good as can be obtained with overcoupled or stagger-tuned i-f circuits if no limiter is used. Receivers for special purposes which utilize narrow system deviations commonly use limiters, and operate on the portion of the i-f response curve which is above 50 per cent of maximum.

It is possible to reduce the frequency deviation, and therefore the band width necessary for the i-f stages, by feeding back some of the demodulated audio output to a reactance tube associated with the superheterodyne oscillator to cause it to follow partially the original frequency deviation. The distortion in the receiver and the overall noise-to-signal ratio are reduced thereby, although the system is rather expensive.

The overall gain in an f-m receiver is usually greater than in an a-m receiver since satisfactory reception may be had at very low levels of input to the antenna terminals, and since, when a limiter is used, the input voltage at its grid terminals must be at least 1 volt. In general, gain is uneconomical in the r-f stages at the frequencies assigned for f-m broadcasting, and practically all the gain is usually obtained at intermediate frequencies. The usual i-f plus converter gains are of the order of 10,000. The band width required lowers the gain per stage, and usually one or two more i-f stages are required in an f-m receiver than in an a-m receiver using the same types of tubes.

The frequency detector directly or indirectly converts the frequency modulation to amplitude modulation and then recovers the audio signal by amplitude detection (see, however, last paragraph of article 8-7). Some means of reducing or preventing response to spurious amplitude modulation due to noise and due to the variation in amplification as the frequency is deviated over the i-f pass band is usually associated with the frequency detector. This may be an amplitude limiter preceding the frequency detector, or the frequency detector itself may be of a type which is non-responsive to amplitude modulation.

A frequency detector with a balanced output, that is, one in which the net rectified output is zero at the mean intermediate frequency, is preferred since spurious audio output can then be produced only during deviation of the frequency due to the desired modulation and is masked considerably by the latter. The d-c output of a balanced frequency detector may be used to control the bias of a reactance tube associated with the oscillator and thereby furnish automatic frequency control. De-emphasis circuits to compensate for the pre-emphasis at the transmitter (corresponding to the voltage across an inductance in series with a resistance when the combination has a time constant of 75 microseconds), and tone-control circuits, are usually included in the a-f system.

Antenna input systems, the superheterodyne oscillator, and the first detector are usually the same in f-m receivers as in a-m receivers for about the same transmitted frequency except that there is more tolerance in tuning and in frequency drift in f-m broadcast receivers than in narrow-band receivers.

Automatic volume control may be incorporated and is desirable to keep the voltage applied to the input of the limiter or non-amplitude-responsive detector at a level which prevents overall response to rapid variations of the net antenna input voltage over as wide a range as possible.

Certain types of f-m broadcast receivers are difficult for a novice to tune by hand since there are multiple tuning positions where there is almost equal volume of response to the desired program. Minimum harmonic distortion is obtained in only the position which corresponds to the most linear portion of the frequency detector characteristic. The program is demodulated in the other positions by the slope of the sides of the i-f response curve and sometimes by the reverse slope of the skirts of the frequency detector characteristic. A greater volume of even-harmonic distortion than fundamental often is produced between the several tuning positions for maximum response. Accurate tuning is also required for the most effective reduction of response to impulse noise.

A zero-center-indicating meter operated by the d-c output of a balanced frequency detector makes an excellent tuning indicator. It indicates zero for the proper tuning position, and the direction of the deflection shows the direction of any mistuning. Twin electron-ray tuning indicators, in which the illuminated portions of the opposite halves are unequal except in the proper position, are sometimes used in broadcast receivers.

Receivers for operation in both the a-m and f-m broadcast bands generally use many of the circuit components in both bands. The intermediate frequency for the f-m section is much higher than for the a-m section, but transformers incorporating tuned circuits for both frequencies are quite satisfactory. The r-f and converter stages tolerate a minimum of switching because of the high frequency and are often separate for the two bands in the more expensive receivers. The audio amplifier and power-output stages are usually common to both sections. Particular care in the design of the audio amplifier and sound reproducer is justified since low harmonic distortion and excellent signal-to-noise ratio are realizable with frequency modulation.

Figure 1 is the circuit diagram of the r-f, i-f, and detector portions of a low-priced fm-am receiver. The desired band is selected by a ganged switch for the r-f, oscillator, converter, avc, and detector circuits. Both bands are tuned with a two-gang condenser having separate stator sections for the two ranges. The f-m section includes a broad-band input transformer for a 300-ohm transmission line, a tuned input circuit to the converter, and delayed avc to obtain the proper input level for the ratio-type frequency detector. The a-m section includes a condenser-tuned low-impedance loop and series loading coil, means

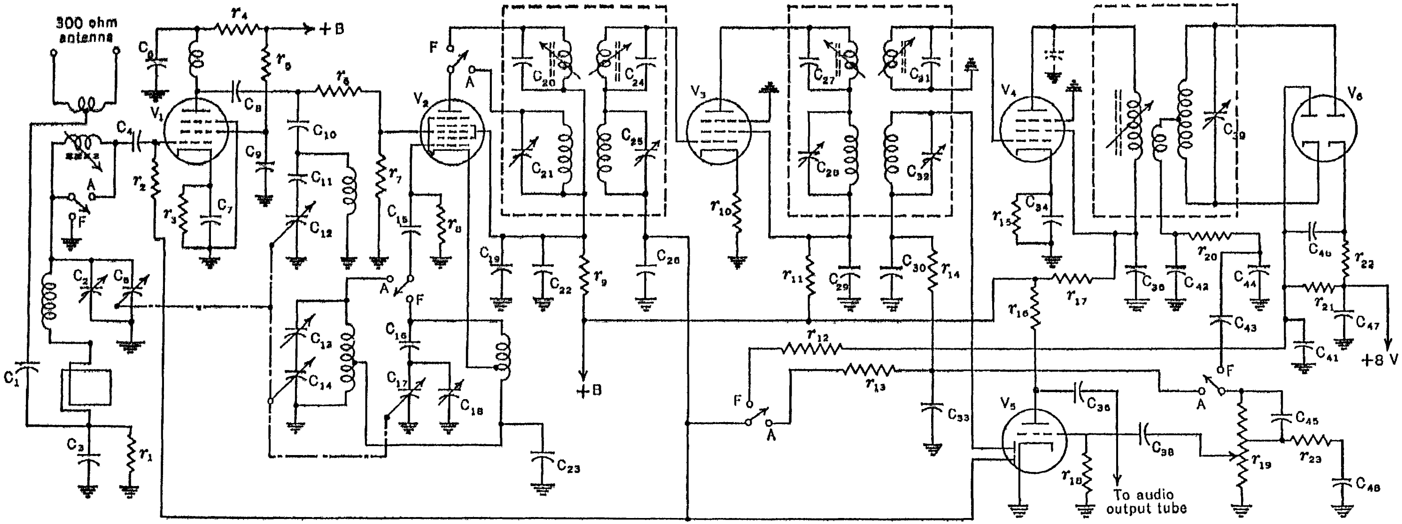


FIG. 1. Circuit of Small FM-AM Superheterodyne (exclusive of output stage and power supply)

Capacitors

2–10 $\mu\mu f$;	C2, C18
10 $\mu\mu f$;	C6
30 $\mu\mu f$;	C10, C11, C16, C19, C20, C24, C27, C31
30–50 $\mu\mu f$;	C39
50 $\mu\mu f$;	C15, C40, C41, C45
100 $\mu\mu f$;	C4, C7, C30, C33, C42, C44
100–200 $\mu\mu f$;	C21, C25, C28, C32
200–500 $\mu\mu f$;	C13
500 $\mu\mu f$;	C23
1000 $\mu\mu f$;	C8, C36
2000 $\mu\mu f$;	C1
5000 $\mu\mu f$;	C3, C38, C47
0.01 μf;	C9, C22, C29, C34, C35
0.02 μf;	C43, C48
0.05 μf;	C26 Gang (a-m); C5, C14
4.0 μf;	C46 Gang (f-m); C12, C17

Resistors

10	ohms; r6, r10
68	ohms; r3, r15
1,000	ohms; r5, r9, r11, r17
4,700	ohms; r4
15,000	ohms; r21, r22
22,000	ohms; r7, r8, r14, r23
47,000	ohms; r1
100,000	ohms; r20
220,000	ohms; r16
470,000	ohms; r12
1,000,000	ohms; r2
2.2	megohms; r13
10	megohms; r18
	Volume; r19

Tubes

6AQ6;	V5
12BA6;	V1, V3, V4
12BE6;	V2
12H6;	V6

for additional antenna input from the leads of the 300-ohm line in parallel, and an untuned input stage for the converter.

Crosstalk between common-channel f-m stations is reduced in receivers incorporating good a-m rejection means to a point where only the modulation of the stronger of two stations is audible. Cross-modulation in the stages of an f-m receiver due to the nonlinearity of tube characteristics is relatively small, whereas in a-m receivers it is one of the major causes of interference at certain input levels.

Phase modulation differs from frequency modulation only in the manner in which the frequency modulation index (deviation ratio) is caused to vary with the modulating frequency. In phase modulation it is proportional to the product of the amplitude and frequency of a modulating component, whereas in frequency modulation it is proportional to the amplitude of the component. Receivers for phase modulation are therefore like receivers for frequency modulation, designed with an i-f band width suitable for the maximum frequency deviation from the center frequency, but with an audio filter following the frequency detector to restore the original amplitude of the modulating component for the output. Pre-emphasis of the higher audio frequencies, as specified by the FCC for f-m broadcasting, gives some of the characteristics of phase modulation to this portion of the transmission.

6. FREQUENCY DETECTORS

Frequency detectors usually consist of some form of i-f slope filter which has a linear variation in output voltage with frequency deviation to the maximum assigned deviation, preferably fed with constant current from the last i-f amplifier, and followed by a conventional amplitude detector. Detectors employing a phase shift between the voltages applied to separate grids of multielectrode tubes corresponding to the frequency deviation have been described but are not commonly used. Circuits which provide frequency detection in a single tube and which are substantially unresponsive to amplitude modulation are in the developmental stage. The detectors described herein have had commercial application.

SIMPLE SLOPE FILTERS. A loosely coupled i-f transformer tuned to one side of the mean intermediate frequency and operated at a point on the side of the resonance curve where the response is about 72 per cent of maximum provides a simple slope filter. The conditions differ with the coupling, but, for example, when the coefficient of coupling between primary and secondary windings is $0.3/Q$, where Q applies to both primary and secondary windings, the variation in amplitude is substantially linear over the range in frequency in which the amplitude varies from 50 to 95 per cent of the maximum for that stage alone. The frequency characteristic for this condition is shown in Fig. 2. The Q of the circuits is reduced by loading so that Qf/F is 0.2, where f is the expected frequency variation from the mean and F is the mean intermediate frequency. The frequency for maximum response for the stage is separated from the mean frequency by $0.35\ F/Q$ and may be either above or below it. Full use of the most linear portion of the characteristic results in 22.5 per cent amplitude modulation, whereas a single tuned circuit provides only 15 per cent. The i-f amplifier is tuned to flatness to operate about one of the operating points mentioned. Final demodulation is obtained in a conventional diode rectifier following the filter.

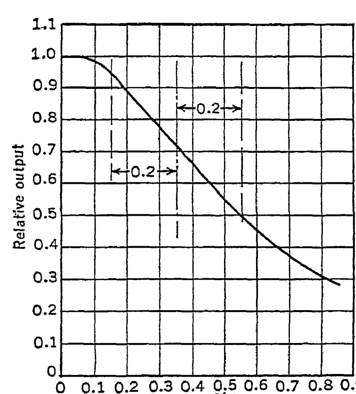


FIG. 2. Frequency Characteristic of Slope Filter

This type of frequency detector is not often used since it provides no inherent balance against spurious amplitude modulation. The output due to impulse noise in this type of frequency detector is much more disturbing than in one which is balanced for zero output at the mean intermediate frequency.

DISCRIMINATORS. Discriminators as shown in Figs. 3(a) or 3(b) are used widely as frequency detectors since they provide an inherent balance against amplitude modulation at the mean intermediate frequency and require only a minimum number of components and adjustments. The i-f voltages applied to the two diodes are respectively the sum and difference of the primary voltage and one-half the secondary voltage of the transformer, which is tuned to the mean intermediate frequency. Since these voltages differ in phase by 90° at the mean frequency, the diode voltages are then equal. The rectified voltages are opposed so that there is no net output at the mean frequency. The voltage applied to one diode increases, while that applied to the other decreases, at higher instantaneous frequencies since the secondary voltage then lags by more than 90°. The reverse is true at lower instantaneous frequencies, and a directional output substantially proportional to the instantaneous frequency deviation over a considerable range is obtained. Thus an audio voltage is available at the output terminals which has an amplitude proportional to the frequency deviation of the applied signal. A d-c output which depends on the difference between the mean applied frequency and the frequency to which the unit is tuned is also developed if these are unequal.

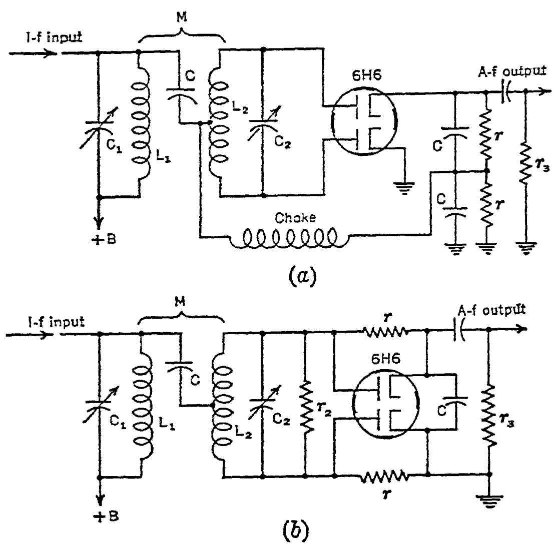


FIG. 3. Discriminator Circuits

Figure 4 illustrates typical shapes of characteristic curves obtained in a discriminator stage when the primary and secondary inductances and Q's are equal. The curves are plotted in terms of Qf/F, where f/F is the ratio of the instantaneous frequency deviation to the center frequency. The actual output is obtained by multiplying the relative output by $\sqrt{2}eQL\omega I$, where e is the rectification efficiency of the diode, L is the inductance of the windings, ω is the mean intermediate angular frequency, and I is the rms value of the current supplied by the last i-f stage. The linearity and the magnitude of the output are near optimum for the conditions shown when the product of the Q and the net coefficient of coupling between windings is about 2.7 although many similar curves may be obtained when the primary and secondary inductances and Q's are unequal.

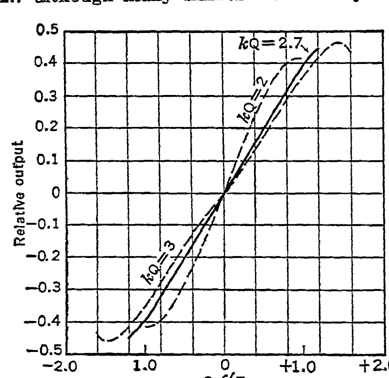


FIG. 4. Discriminator Characteristics

It is important in obtaining a symmetrical characteristic curve that the secondary winding and circuit be symmetrical with respect to ground. This requires that the center tap on the secondary be placed properly, that the coupling between the halves of the secondary be close, and that the leads to the diodes be short and have no spurious couplings to other parts of the circuit. The diode capacitances should be equal since that of one aids, while that of the other opposes, the magnetic coupling between the windings. The required Q, which includes the effects of diode loading, and which may be reduced by circuit loading if necessary to obtain the proper band width, may be estimated from the abscissas of Fig. 4 over which the linearity is satisfactory for the expected ratio of f/F. The coupling and by-pass condensers C are made only large enough to have low impedance at the intermediate frequency while their impedance at audio frequencies is large. The current through each diode load resistance r should be only that rectified by the individual diode. Any shunt resistance across the output terminals, such as r_3, is made large so that the current through it and the two resistors r in series will not bias appreciably the diode delivering the smaller instantaneous output.

The i-f choke shown in Fig. 3(a), which is effectively in parallel with the primary inductance, may be omitted if the diode loads are connected as in Fig. 3(b). This is permissible if the desired Q can be obtained with the increased loading effect which the diodes present for this connection. Each individual diode presents a load of $r/1.8$ in Fig. 3(a) and $r/2.8$ in Fig. 3(b), where r is as indicated in the figures, and the diode rectification efficiency has the usual value of about 90 per cent. The desired Q may be obtained either by loading the windings, as with r_2, or by designing the individual windings with the diameter and spacing of turns which result in the required value.

The frequency at which the response is zero is adjusted by tuning the secondary circuit of the discriminator. The symmetry of the response curve about the zero point is adjusted by tuning the primary circuit. This procedure is easier than adjusting each individual circuit for resonance at the center frequency.

Some special-purpose receivers are arranged to respond to either frequency modulation or amplitude modulation in the same transmission band by providing a reversing switch for one of the diodes of the discriminator. When one is reversed from the polarity indicated in Fig. 3, the sum, rather than the difference, of the rectified voltages is applied to the output, which is proper for the demodulation of a-m waves.

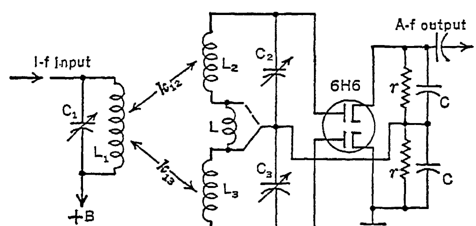


Fig. 5. Frequency Detector with Side-tuned Circuits

SIDE-TUNED CIRCUITS AS FREQUENCY DETECTORS. Another frequency detector which has zero output at the mean intermediate frequency consists of separate tuned circuits C_2L_2 and C_3L_3 individually tuned slightly above and below the center frequency and connected as shown in Fig. 5. The response curve is shown in Fig. 6. The stage is tuned by adjusting the peaks of the S curve, each of which depends chiefly on the tuning of one of these circuits but is influenced slightly by the tuning of the other. An alternative tuning method is to transfer the connection between the two tuned circuits from the full-line to the dotted-line connection after tuning both circuits to the center frequency. The incremental inductance L is chosen to shift the resonance by the proper amount in opposite directions.

Neglecting the coupling between the two side circuits caused by the common primary circuit C_1L_1, to which they are individually coupled, the best linearity of the characteristic is obtained when $f_1/F = 0.75/Q$, where f_1 is the frequency displacement of either resonant side circuit from the center frequency. The relative output for this case is shown in Fig. 6 plotted against relative instantaneous frequency deviation Qf/F. The actual output depends on the coupling to circuit C_1L_1. The capacitance C_1 of the driving tube may be tuned out at the center frequency, and the effects of the primary circuit over the band may be minimized by loading it with resistance. The presence of the tuned primary circuit slightly improves the linearity of the S curve near its peaks by an amount not shown in Fig. 6, and depending on its coupling with the side circuits. To obtain the best linearity for both positive and negative deviations it may be necessary to use slightly different Q's in the two side circuits and slightly different coefficients of coupling between the primary and these circuits. Such expedients compensate for the fact that the resonant peaks of the side circuits are not at the same frequency.

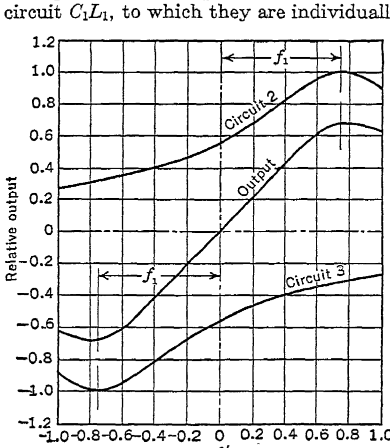


Fig. 6. Frequency Characteristic with Side-tuned Circuits

RATIO-TYPE FREQUENCY DETECTORS. One type of frequency detector which can be made to be non-responsive to any undesired amplitude modulation has been called a ratio detector, since the net output is approximately proportional to the ratio of the i-f voltages applied to the two diodes, although the process by which this is accomplished is indirect, and a more complete analysis shows that the ratio of the circuit impedances rather than the voltages is involved. Two of many possible arrangements are illustrated in Figs. 7(*a*) and 7(*b*). In Fig. 7(*a*) the filter is quite similar to the discriminator previously described except that the components are designed to tolerate a greater a-c load current. In Fig. 7(*b*) the side-tuned circuits C_2L_2 and C_3L_3 are in series with no magnetic coupling to the choke feed L_1, while capacitance C_1 is as small as possible to reduce the undesired coupling between these circuits. In both arrangements condensers C are only large enough to have low impedance at the intermediate frequency.

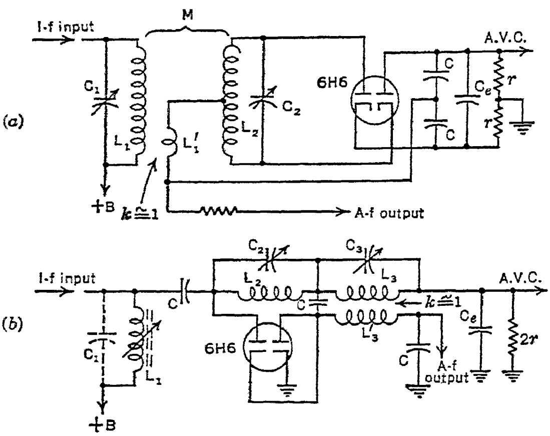


FIG. 7. Ratio-type Frequency Detector Circuits

The diodes are in series and are therefore forced to carry the same rectified load current which charges a large electrolytic condenser C_e to a voltage which depends on the mean i-f signal level. During desired frequency modulation or undesired amplitude modulation, this voltage, all or a part of which may be used also for automatic volume control, remains constant and establishes conditions for the mean rectification efficiency of the diodes and for a voltage drop in the a-c impedance of the source. The action is illustrated by the diagrams of Fig. 8 which are not to scale but show, diagrammatically, superimposed regulation curves for the diodes (one inverted to show the division of the electrolytic condenser voltage E) for several conditions. Figure 8(*a*) shows the relations between the output current and the component voltages which lead to operation at a particular value of output current at one input level and frequency. The peak voltages applied to the input terminals of the diodes are E_1 and E_2 respectively. They have maximum values for no rectified current since the a-c diode input current is from 1.67 to 2.0 times the rectified current, depending on the rectification efficiency of the diodes. The rectified voltages, whose sum is E, are the input voltages multiplied by the individual rectification efficiencies and are e_1 and e_2 respectively. The diodes are in series, and the common-current operating point is shown by the marked intersection.

The a-c diode input current is in phase with the input voltage, but the filter networks carry substantial reactive components of current. Furthermore, the relations between the diode input and output currents and voltages are non-linear. Therefore the characteristics cannot be expressed readily in terms of the circuit constants. However, each circuit and associated diode has a definite regulation curve for each frequency and for each current level supplied by the last i-f tube. The effective impedance which determines the output voltage drop in terms of the output current depends primarily on the impedance of the tuned circuits and secondarily on non-linear functions of the current and voltage.

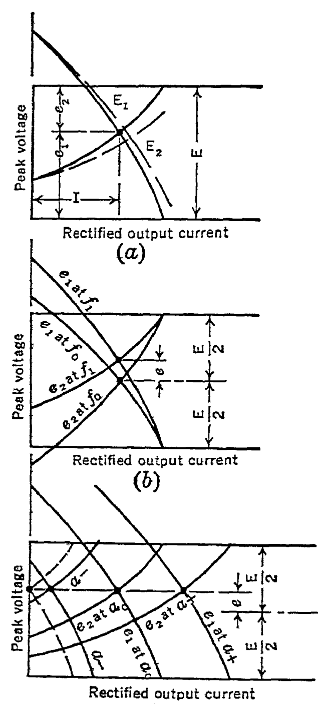


FIG. 8. Regulation Curves in Ratio-type Frequency Detectors

Figure 8(*b*) shows the effect of deviating the frequency. This changes the filter impedances and consequently the effective output impedance and the slope of the regulation curves. The difference between either diode output voltage and its mean value, $E/2$, varies at the modulating frequency and represents the desired audio output.

Figure 8(*c*) shows the effect of amplitude modulation of the input signal at a given frequency. The voltage E remains constant at its mean value by virtue of the charge on the electrolytic condenser. The output current varies instantaneously with amplitude modulation although its mean value is substantially constant. If the degree of inward modulation is sufficient to reduce the rectified current to zero as illustrated by the dotted lines, rectifying action is lost and intolerable distortion results. The resistance r is made smaller than in conventional discriminators to allow for a margin in inward modulation without reaching these limits. The ratio of the uniform signal to the instantaneous signal which will produce cut-off is the overdrive and should be at least 2 to 1. It is particularly important that the relations be so chosen that the detector is not cut off when the instantaneous voltage swings to its lowest value as the frequency is deviated over the selectivity response curve of the receiver.

It can be shown that with ideal diodes with rectification efficiencies of 100 per cent, and with non-reactive source impedances, the output voltage e is independent of amplitude modulation when the short-circuit currents I_s, as shown by the regulation curves, are the same for the two diodes. The output is then $e = 0.5E(Z_1 - Z_2)/(Z_1 + Z_2)$, where Z_1 and Z_2 are the impedances of the two filter sections. In actual ratio-type frequency detectors the phase angle of the filter impedance varies with the frequency deviation and the diode efficiency depends on both applied voltage and rectified current. It is therefore difficult to compensate perfectly for amplitude modulation over a wide range of either signal level or deviation. The best combination of impedances is determined by trial. For the circuit of Fig. 7(*a*), for example, optimum values of coil inductance L_2 and L_1', mutual inductance m, resistance r, and capacitance C will be found for a particular type of tube and range of operating levels for the best linearity of desired output and reduction of amplitude modulation.

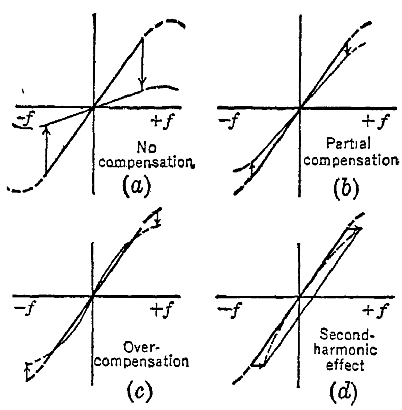


Fig. 9. Amplitude Compensation in Ratio-type Frequency Detectors

Some of the possible detector characteristics for simultaneous amplitude and frequency modulation are illustrated in Fig. 9. The curves drawn with heavy lines show the output during maximum outward amplitude modulation. Figure 9(*a*) is for a conventional discriminator without the compensating effect of the ratio-type detector. Figures 9(*b*) and 9(*c*) are for partial and overcompensation respectively. In (*c*) high input level produces less output over a portion of the deviation range than lower input.

The diodes carry components of current at the second-harmonic frequency as well as at the fundamental frequency. This current returns through the filter sections and produces a small second-harmonic voltage which shifts the effective phase of the peak voltage to be rectified. This effectively detunes the filter sections synchronously with the amplitude modulation and accounts for an unbalanced characteristic as illustrated in Fig. 9(*d*) between two signal levels. When the variation in input level is due to deviating over a non-uniform selectivity response curve of the receiver, the resulting characteristic may be as shown by the dotted line.

The demodulated signal rises and falls with the applied signal when it is varied slowly, as in tuning. This is an advantage, since the proper tuning position is then indicated by maximum volume.

Since a large degree of reduction of amplitude modulation is obtained in the ratio-type detector stage itself, limiting in the previous stages is not always required, and the signal level at the last i-f tube need not be as high as in receivers using limiters. Automatic volume control may provide sufficient control of signal level.

Multipath transmission through space of the signal applied to the antenna terminals of the receiver may result in amplitude modulation sufficient to reduce the instantaneous input to the diodes in a ratio-type detector below their cut-off level, and in this case a receiver having high i-f gain followed by limiter is superior.

7. LIMITERS

An ideal limiter or limiting system for operation at the intermediate frequency of an f-m receiver delivers an rms output which is independent of the input when the input is above a threshold level. (However, a proposed type "dynamic limiter" gives output proportional to the *average* input but wipes off any a-f amplitude modulation.) It should operate instantaneously and therefore should not include time-constant circuits which delay its recovery after being subjected to a high input voltage, as, for example, a burst of impulse noise. The loading effect of the limiter on associated tuned circuits should not change with signal level.

The greatest value of a limiter is in reducing amplitude modulation, synchronous with the desired frequency modulation, which may be introduced by the deviation of the frequency over symmetrical but slightly rounded i-f response curves. It is also of value in reducing random or impulse noise which occurs while the carrier is deviated from the mean.

GRID-BIAS LIMITERS. Limiters in which the operation is controlled by the bias developed by grid rectification are more commonly used than other types. The fundamental component of the plate current of a grid-bias limiter is maintained substantially constant over a considerable range of input (usually about 10 to 1) by a proper correlation of the bias developed at these levels with the bias necessary to cut off the instantaneous plate current, thereby compensating for the change in angle in each cycle over which the plate circuit is conducting. This relation is most easily obtained in a pentode tube since its plate current is nearly independent of the plate load. A typical circuit is shown in Fig. 10. The output is applied to a circuit tuned to the fundamental frequency (usually the discriminator), and the harmonics are filtered out. A sharp-cutoff tube is used since limiting may then be obtained at low input levels.

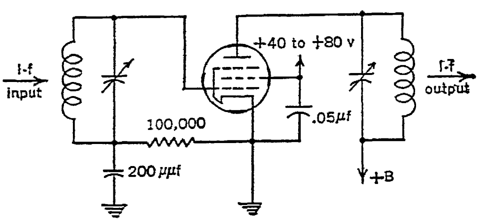


FIG. 10. Grid-bias Limiter

A typical grid-bias limiter static characteristic is shown by the solid line in Fig. 11. If the grid resistor is too small, the efficiency of grid rectification and the developed bias are too small, so that the angle of plate conduction is not reduced sufficiently at high input levels, and the output current rises, as shown by the upper curve. Conversely, if the grid resistor is too large, the output falls as shown by the lower curve. At very high input levels the proper relations cannot be held for any proportions and the output again rises. Although the static characteristic of a grid-bias limiter may be made flat over a wider range than shown by placing resistors by-passed to ground in the plate- or screen-supply circuits, the overall operation in the presence of impulse noise is not satisfactory since the conditions following a burst of noise are not normal and the output suffers during the recovery time of the plate or screen circuits.

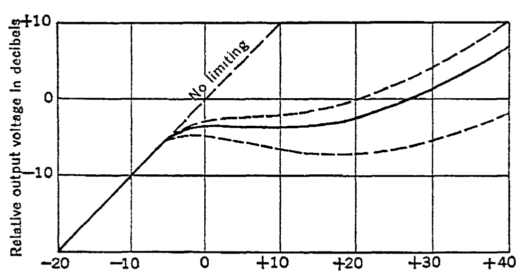


FIG. 11. Grid-bias Limiter Characteristics

An approximate rule for flat limiting which holds for pentode tubes may be used if the coefficient a in the expression $i_g = ae^{3/2}$ for the grid current i_g in terms of the instantaneous applied grid voltage e is known. The product $arE_c^{1/2}$ should be 35 or 40, where r is the grid resistor in ohms and E_c is the d-c grid voltage necessary to cut off the plate current with the screen voltage selected. The rms plate current is then approximately $0.5g_mE_c$ in the range of rms input voltage from 0.7 to $7.0E_c$, where g_m is the transconductance of the tube with small negative bias at the screen voltage selected. The level at which limiting occurs with common tubes is ordinarily between 1 and 3 volts but may be adjusted over a small range by selecting the screen voltage.

Cascaded limiters are often used to cover a greater range of input signals over which limiting is effective. The voltage applied to the grid of the last limiter from that developed in the tuned output circuit of the previous limiter is made to fall at a point below the final upward curvature of the static characteristic so that an additional 10-to-1 range in level may be handled. All the i-f tubes may act as limiters when the i-f amplifier is used only for f-m signals. The rectified d-c voltage in the grid resistor may be filtered and used for avc bias.

The time constant of the r-C combination in the input circuit of a grid-bias limiter should be as short as is consistent with little loss of i-f gain. The recovery time is then fast enough so that the program is not eliminated for an audible interval after a burst of impulse noise.

PLATE LIMITER. A plate limiter operates when the plate voltage of a triode or pentode swings downward to that portion of the plate-current plate-voltage characteristic where an increase in grid voltage in the positive direction produces no increase in plate current. The stage is operated with very low plate-supply voltage and with a high-impedance plate load. During negative grid voltage swings the plate current is cut off. The instantaneous output current swings between the maximum and zero, and tends to deliver a rectangular wave at high input levels. The rms output voltage increases slightly with increase of input level until the output current has assumed the rectangular wave form, and is then somewhat less than half the plate-supply voltage. The harmonics of voltage in the output are eliminated in the tuned circuit.

Heavy grid current loads the input circuit of a plate limiter during positive grid swings. This may be restricted by a resistor in series with the grid lead but may still influence the selectivity and gain of the input tuned circuit. A plate limiter has the advantage of rapid recovery time if the plate- and screen-supply voltages are not influenced by the tube load but is more often used in clipping and shaping pulses than in amplitude limiting in f-m receivers.

LOCKED-IN OSCILLATOR. An oscillator operating at the intermediate frequency or some submultiple thereof may be used to cause the receiver to be non-responsive to amplitude modulation and may be synchronized or locked in by the i-f signal and then follow its deviation. An oscillator remains in synchronism over a wider band, and its output is slightly greater, for high signal inputs than for low. It has been found to be most satisfactory when operated at a submultiple of the intermediate frequency. It then has the advantage of having an output frequency which will not feed back to the previous i-f stages. However, it requires an input signal above 1 volt and in this respect is no more satisfactory than a grid limiter. Discriminators for use with synchronized oscillators must be specially proportioned to take care of the interaction with and the loading of the oscillator circuits.

A frequency detector which operates as a locked oscillator in which the frequency is controlled over the deviation range by quadrature feedback from the plate circuit of a heptode tube, and which simultaneously provides an audio output in the plate circuit, has been described. The output is linearly proportional to the frequency deviation and is independent of i-f amplitude, provided the latter is great enough to maintain synchronism.

BIBLIOGRAPHY

Arguimbau, Discriminator Linearity, *Electronics*, Vol. 18, 142–146 (March 1945).

Armstrong, E. H., A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation, *Proc. I.R.E.*, Vol. 24, 689–740 (May 1936).

Beers, G. L., A Frequency-dividing Locked-in Oscillator Frequency-modulation Receiver, *Proc. I.R.E.*, Vol. 32, 730–737 (December 1944).

Bradley, W. E., Single-stage F-m Detector, *Electronics*, 88–91 (October 1946).

Carnahan, C. S., and Kalmus, H., Synchronized Oscillators as F-m Receiver Limiters, *Electronics*, Vol. 17, 108–111, 332–342 (August 1944).

Crosby, M. G., Frequency Modulation Noise Characteristics, *Proc. I.R.E.*, Vol. 25, 472–514 (April 1937).

Meyers, S. T., Non-linearity in Frequency-modulation Radio Systems Due to Multipath Propagation, *Proc. I.R.E.*, Vol. 34, 256–265 (May 1946).

Roder, H., Amplitude, Phase and Frequency Modulation, *Proc. I.R.E.*, Vol. 19, 2145–2176 (December 1931).

Smith, D. B., and Bradley, W. E., The Theory of Impulse Noise in Ideal Frequency-modulation Receivers, *Proc. I.R.E.*, Vol. 34, 743–751 (October 1946).

Wheeler, H. A., Common-channel Interference between Two Frequency-modulated Signals, *Proc. I.R.E.*, Vol. 30, 34–50 (January 1942).

DISTORTION AND INTERFERENCE IN F-M SYSTEMS

By B. D. Loughlin

In f-m systems, just as in a-m systems, distortion can result from a non-linearity of the input-output characteristic of the modulator or frequency detector. However, the modulator of the transmitter can generally be properly designed to have a substantially linear input-output characteristic over the operating range. In the commonly used phase modulators which are inherently non-linear for large phase deviations, satisfactory linearity is obtained by restricting the operating range to use small phase deviations at the modulator, followed by frequency multiplication to obtain the desired frequency modulation. It is also relatively straightforward to obtain linear frequency detection from the commonly used f-m detectors when they are receiving an ideal f-m signal of constant amplitude. Thus the commonly used frequency modulators and detectors are generally designed so that they contribute only a small amount of distortion to the f-m system.

This section treats the special forms of distortion which are unique to an f-m system and which arise from translating the f-m signal through the amplifiers of the transmitter and receiver and through the transmission medium. When the f-m signal passes through the amplifiers of the transmitter or receiver, f-m distortion can result because of inadequate band width or non-linear phase characteristic. Also, if the transmission characteristics of the amplifiers are not flat over the frequency deviation of the applied signal, spurious amplitude modulation synchronous with the frequency modulation is introduced, which may cause distortion if there is incomplete rejection of amplitude modulation by the f-m detector system. Another serious form of f-m distortion, producing both spurious amplitude and frequency modulation, results from multipath transmission between the transmitter and receiver.

8. F-M DISTORTION FROM NON-UNIFORM AMPLITUDE AND PHASE CHARACTERISTICS

When an f-m signal is translated through an amplifier or network having a non-uniform amplitude or phase characteristic, some spurious frequency modulation (in other words, f-m distortion) results. The f-m distortion results because the various sideband components of the f-m signal are translated with different amplitude and delay and thus do not correctly combine in the output. The first approximation to the distortion can be obtained by a quasi-steady-state analysis. In a quasi-steady-state analysis, it is assumed that at any instant the f-m signal is translated with an amplification and delay determined by the steady-state amplitude and delay characteristics measured at a frequency equal to the applied instantaneous frequency. The delay used is, of course, the envelope delay, that is, the slope of the phase characteristic at the particular frequency.

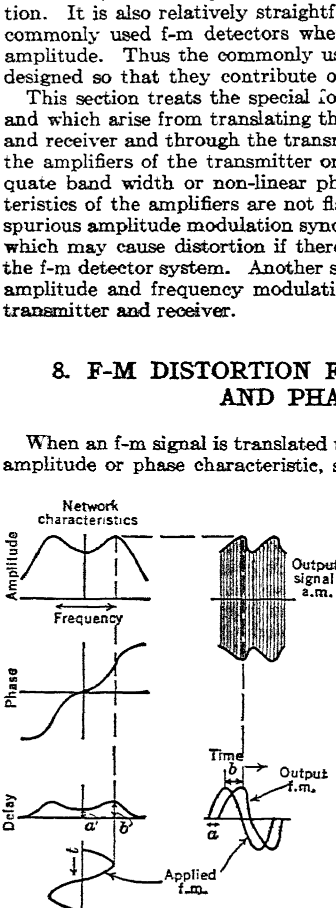


FIG. 1. Quasi-steady-state Approximation for F-M Distortion

To illustrate the quasi-steady-state solution, assumed amplitude and phase characteristics, with the resulting delay characteristic, are shown in Fig. 1. An applied f-m signal with a sine wave of modulation is shown together with the resulting output signal instantaneous frequency and amplitude modulation. It can be seen that the non-uniform amplitude characteristic introduces spurious amplitude modulation and no f-m distortion but that the non-uniform delay results in an f-m distortion due to different delay for different parts of the audio cycle. The resulting distortion can be found by a graphical Fourier series analysis, or a Fourier series expansion of the output f-m equation. In accordance with this approximate analysis method the output frequency modulation would be:

$$\text{f.m.} = a \sin p(t - t_d) \tag{1}$$

where a = maximum frequency deviation, p = angular modulation frequency, and t_d = delay of circuit (a function of instantaneous frequency). In this $t_d = t_0 + F(a \sin pt)$, Practical matters, such as ease of alignment, tolerance of manufacture, and adequate selectivity, frequently dictate the use of undercoupled circuits in commercial f-m receivers, thus resulting in a non-uniform amplitude characteristic. Even receivers designed to have a flat selectivity curve are frequently in trouble as the result of spurious amplitude modulation when the receiver is not accurately tuned or when the set drifts out of alignment. Thus it is desirable that the f-m detector system have fairly complete rejection of amplitude modulation in order to reduce distortion.

The amount and type of harmonic distortion produced by the spurious amplitude modulation is determined by the manner of the detector response to it. As indicated by Figs. 9(*b*) and 9(*d*) (p. 8-23), the response of the detector to amplitude modulation may be in a balanced or an unbalanced manner, or a combination of the two. As shown here, by Fig. 3, a receiver having a rounded-top selectivity curve, and correctly tuned, gives predominantly third-harmonic distortion when the detector has a balanced response to

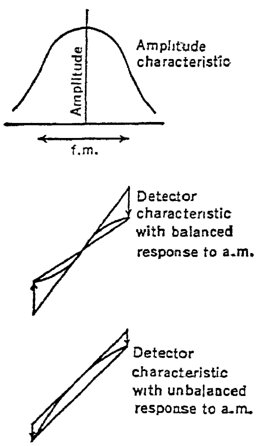


FIG. 3. Distortion from Amplitude Modulation

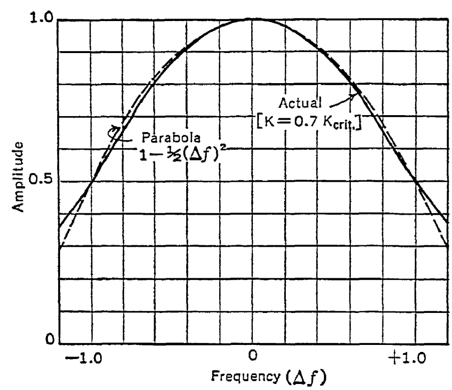


FIG. 4. Amplitude Characteristic for Two Cascade Double-tuned Transformers

amplitude modulation and gives predominantly second-harmonic distortion when the detector has an unbalanced response to it.

The harmonic distortion can be calculated by writing the equation for the f-m detector input-output characteristic in terms of both input signal frequency and amplitude. Then, the equation for the signal amplitude vs. the signal frequency is found from the selectivity curve. By considering the input signal to be frequency-modulated by a sine wave, the resulting output from the detector can be written as a trigonometric series. Substitution of suitable trigonometric expansions will give the fundamental and harmonic output signals.

As an example, consider a receiver with an i-f amplifier system including two coupled circuit transformers having 0.7 of critical coupling and a band width such that the response is down 6db at full system deviation (at ±75 kc for broadcast frequency modulation). Figure 4 shows that the amplitude characteristic of such an i-f system can be closely approximated by a parabola, giving: $A_1 = 1 - {}^1/_2\,(\Delta f)^2$, where $\Delta f = 1$ corresponds to full system deviation. Now the response of a balanced discriminator can be represented as $E_0 = A_2 k(\Delta f)$, where A_2 is the applied signal amplitude, and k relates to the f-m detector slope. If the amplitude modulation of the signal applied to the detector is effectively reduced by some a-m reduction factor (a), then the amplitude modulation due to the i-f selectivity is effectively reduced to $A_2 = 1 - \frac{a}{2}(\Delta f)^2$. Then the detector output is:

$$E_0 = \left[1 - \frac{a}{2}(\Delta f)^2\right] k\,(\Delta f) = k\,(\Delta f) - \frac{a}{2} k\,(\Delta f)^3 \tag{3}$$

Applying a sine wave of frequency modulation, $\Delta f = \sin pt$ (for 100 per cent modulation), then $E_0 = k \sin pt - \frac{a}{2} k \sin^3 pt = k\left(1 - \frac{3}{8}a\right)\sin pt + \frac{ak}{8}\sin 3pt$, thus giving: Per cent third-harmonic distortion $= \dfrac{100a}{8(1 - {}^3/_8\,a)}$. For an a-m reduction factor (a) of 0.35, this relation can be seen to give approximately 5 per cent of third-harmonic distortion.

It can be seen from the above that the exact amount and the harmonic order of the distortion will be affected not only by the selectivity and the manner and degree of the f-m detector response to amplitude modulation but also by the alignment of the detector relative to the intermediate frequency and by the tuning of the center frequency of the f-m signal relative to the center frequency of the intermediate frequency. In particular, it is possible to obtain regions of high audio output and large distortion when, as the signal is detuned, the carrier level at the detector system falls below that necessary for good a-m rejection and when the carrier, at the same time, is tuned on the steep side-slopes of the i-f response. Such distorted side responses can be considerably reduced by using a rounded-top i-f selectivity in conjunction with a well-designed ratio f-m detector.

10. DISTORTION DUE TO MULTIPATH RECEPTION

When the same radio signal is received over several paths having different delay times, the several signals may combine to give increased or decreased amplitude and/or an advance or delay in the resulting carrier phase. For fixed differences in time delay of the paths, the relative phase of the signals will vary with frequency of the signal. Thus multipath reception of an f-m signal will result in spurious amplitude and spurious phase modulation relative to the desired signal. The various signals can combine so that a substantial null in transmission exists at certain frequencies. As the carrier deviates through such a null frequency, a sudden downward amplitude modulation results in conjunction with a rapid change in phase. The sudden change in phase can result in significant spurious frequency modulation. It appears that, where multipath transmission is expected, it is of primary importance that the f-m detector system have good a-m rejection, particularly in terms of rapidity of action and amount of downward amplitude modulation that can be accommodated. In general, the resulting f-m distortion cannot be eliminated; however, a large amount of amplitude modulation is generally produced before such distortion is severe. Thus, good a-m rejection helps considerably but cannot eliminate multipath distortion effects.

Although numerous examples of multipath transmission distortion have been cited on the higher frequencies of 50 to 100 Mc,* it does not appear to represent a serious threat to the f-m broadcast industry. However, multipath transmission distortion virtually makes voice communication by frequency modulation impractical on the long-distance short-wave bands of 5 to 30 Mc.† The many paths of transmission occurring during "skip" transmission on these frequencies result in the familiar selective fading frequently producing serious f-m distortion, particularly as the deviation is increased.

11. CROSS-TALK AND BEATNOTE INTERFERENCE

Interference in f-m systems may arise from other generated signals, such as communication f-m or a-m signals, either received directly or through spurious receiver responses, or it may arise from noise signals of such form as fluctuation noise or impulse noise. The response of an f-m system to such interferences is, in general, quite different from that of an a-m system. For example, cross-modulation on amplitude modulation, where the modulation of a strong undesired a-m signal produces amplitude modulation of a desired signal, has no exact equivalent in an f-m system. The non-linearities which produce cross-modulation in amplitude modulation produce some spurious signals in an f-m system which can have the frequency modulation of both signals, but no direct cross-modulation of one carrier modulation upon the other carrier is produced.‡

If two carrier signals of different frequency exist in a linear system, the resulting signal (see Fig. 5) has amplitude and phase modulation at the difference frequency rate. When the ratio of the two signals is substantially different from unity, the fractional amplitude modulation and the radian phase modulation are equal to the fractional signal ratio, that is, the ratio of the weaker signal amplitude to the stronger signal amplitude. The average frequency of the resulting signal, being the average number of cycles per second, is the frequency of the stronger signal. Since the instantaneous frequency modulation is determined by taking the differential of the phase modulation, the resulting signal has a frequency modulation that is not only proportional to the signal ratio but also directly

* Frequency-Modulation Distortion Caused by Multipath Transmission, M. S. Corrington, *Proc. I.R.E.*, Vol. 33, 878 (December 1945).

† Observations of Frequency-modulation Propagation on 26 Megacycles, M. G. Crosby, *Proc. I.R.E.*, Vol. 29, 398 (July 1941).

‡ Two Signal Cross-modulation in a Frequency-modulation Receiver, H. A. Wheeler, *Proc. I.R.E.*, Vol. 28, 537 (December 1940).

proportional to the frequency difference between the carriers. Thus, if the two carriers are applied to an ideal f-m detector system, the average detector output will be determined by the average frequency of the stronger signal and an f-m beatnote will occur in the detector output having a frequency equal to, and an amplitude proportional to, the difference in frequency between the carriers.

When the ratio of the two beating signals is far from unity, the a-m, p-m, and f-m beatnotes are all substantially sinusoidal. However, as the ratio approaches unity, the beatnote wave forms depart from sinusoidal, as shown in Fig. 6. In particular, the f-m beatnote approaches a pulse wave form as the signal ratio approaches unity.

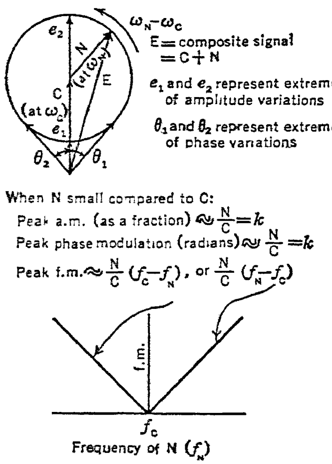


FIG. 5. Addition of Two Carrier Signals

In an ideal f-m detector system, the detector responds only to the frequency modulation of the stronger signal, resulting in substantially no cross-talk from a weaker signal.* This applies to co-channel as well as adjacent channel signals, if the desired-to-undesired-signal ratio is measured just preceding the circuit which effectively reduces or which is effectively non-responsive to the amplitude modulation of the signal. For co-channel interference, an f-m beatnote, of variable frequency, results which has considerably less amplitude within the audio spectrum (owing to the triangular f-m beatnote spectrum as shown in Fig. 5) than would exist in an a-m system. The direct audible beatnote interference does not exist for adjacent-channel interference in a conventional f-m receiver.

In practical f-m receivers the ideal performance in regard to co-channel and adjacent-channel interference is not realized because of inadequate a-m rejection. The amount by which the signal ratio can approach unity is generally limited by the downward a-m rejection capability of the system. This generally means that the signal ratio can get to within only 3 to 6 db of equality before cross-talk results, even with good f-m detectors. When considering the signal-to-interference ratio for the adjacent channel case, the most adverse condition during modulation must be taken. This exists when the undesired signal has maximum deviation toward the center of the i-f pass band, and, owing to the sharp skirt selectivity, this condition may differ considerably from the unmodulated case.

Another practical limitation results from inadequate rejection of superaudible amplitude modulation produced in adjacent-channel interference, where the beatnote generally exceeds 200 kc. The grid-bias limiter plus balanced discriminator type of f-m detector system suffers from this limitation. The usual limiter grid-circuit and diode load time-constants do not permit following of the superaudible amplitude modulation, resulting in an effective unbalancing of the discriminator and a change in average output of the limiter. This limitation frequently requires a signal-to-interference ratio of about 20 db to eliminate cross-talk from adjacent channel signals, when this f-m detector system is used. Detector systems not including the equivalent of this time-constant limitation are generally able to tolerate a 3- to 6-db signal-to-interference ratio to eliminate adjacent channel cross-talk.

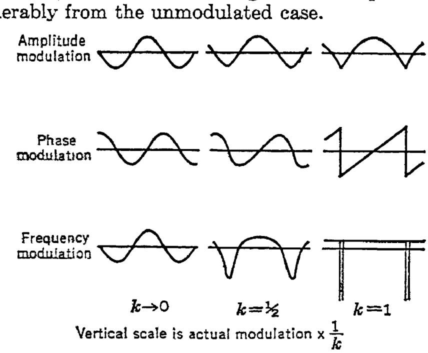


FIG. 6. Beatnote Wave Forms

12. FLUCTUATION NOISE INTERFERENCE

Fluctuation noise, such as thermal noise of resistive impedances, and shot noise and divisional noise of vacuum tubes, can be considered equivalent to a uniform spectrum of

* Common Channel Interference between Two Frequency-modulated Signals, H. A. Wheeler, *Proc. I.R.E.*, Vol. 30, 34 (January 1942).

energy in which the components have random phase or timing. When the noise is small compared to the signal, any individual noise component will beat with the carrier to give an f-m beatnote as illustrated in Fig. 5. Thus the resulting audio noise consists of predominantly high audio-frequency components giving a characteristic high-frequency hiss for f-m noise as compared to the uniform spectrum with considerable low-frequency rumble for a-m noise.*

To find the rms value of the audio noise, the output noise spectrum can be squared, the resulting squared spectrum can be integrated over the audio band, and the square root of the integral taken. Using this for the simple case of an audio system with uniform response and sharp cut-off, the f-m signal-to-noise ratio to the a-m signal-to-noise ratio (called the f-m improvement ratio) is found to be $\sqrt{3}\, f_d/f_a$, when f_a is the maximum audio frequency and f_d is the maximum frequency deviation of the system. For 15-kc audio and 75-kc deviation this gives an f-m improvement of 18.8 db.

By including a de-emphasis low-pass filter in the receiver, which is compensated for by a complementary pre-emphasis circuit at the transmitter, the f-m signal-to-noise ratio can be further improved. In this case the f-m signal-to-noise ratio including de-emphasis to the a-m signal-to-noise ratio not including de-emphasis is given by:

$$\frac{f_d}{f_0\sqrt{1 - \dfrac{f_0}{f_a}\tan^{-1}\dfrac{f_a}{f_0}}}$$

where f_0 is the frequency for 3-db attenuation of the de-emphasis filter. For broadcast frequency modulation with 15-kc audio, 75-kc deviation, and 75-microsecond de-emphasis time constant ($f_0 = 2.12$ kc), this gives an f-m improvement of 32 db.

The above relations are derived on the assumption that the noise is small compared to the carrier signal. In the region where the peak noise is almost equal to the peak carrier, the simple relations are inadequate, and actually the f-m improvement is rapidly lost as the signal is made weaker. The approximate threshold for f-m improvement is when the peak carrier equals the peak noise.

Signal-to-noise ratios are frequently expressed in terms of the rms audio signal output for 30 per cent frequency modulation to the rms audio noise when the carrier is unmodulated. Using this definition and applying the approximate equations at the f-m improvement threshold (where peak carrier equals peak noise after selectivity), the signal-to-noise ratio for broadcast frequency modulation with 150-kc i-f pass band is about 40 db at the threshold. Thus, in this case, the frequently used 30-db signal-to-noise ratio is near the knee of the improvement threshold, and when the receiver has good a-m rejection it is determined by the signal level which approximately gives peak carrier equal to peak noise. For broadcast f-m receivers with 150-kc i-f pass band, a 300-ohm antenna, and an assumed receiver noise factor of 6 db, the threshold level (peak carrier = peak noise) is at about 104 db below 1 volt. This would represent a very well-designed set. Normal design receivers have an improvement threshold around 90 to 100 db below 1 volt.

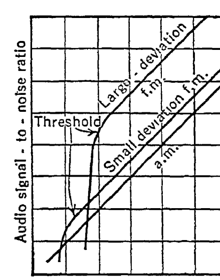


FIG. 7. Small- vs. Large-deviation Ratio FM

When the peak carrier greatly exceeds the peak noise, the signal-to-fluctuation-noise ratio of an f-m system is improved by using a larger deviation ratio. However, a small-deviation-ratio system can have narrow receiver selectivity resulting in less total peak noise and thus a lower threshold signal level, as illustrated in Fig. 7. Thus, entertainment f-m systems where signal-to-noise ratio is important are built using a large deviation ratio, while communication networks where range of coverage is important use a small deviation ratio.

13. IMPULSE NOISE INTERFERENCE

When an impulse, such as automobile ignition interference, is applied to a receiver, a transient carrier pulse results having a duration determined by the band width of the receiver and a frequency determined by the center frequency of the i-f selectors. When this transient is added to a desired carrier, an amplitude and phase modulation of the desired carrier results, depending upon the relative amplitude, frequency, and phase of the carrier and the transient. If the carrier amplitude is larger than the peak amplitude

* Frequency-modulation Noise Characteristics, M. G. Crosby, *Proc. I.R.E.*, Vol. 25, 472 (April 1937).

of the transient, the maximum phase modulation that can result is a pulse of less than 1 radian. In this case the audible interference produced, particularly in large-deviation-ratio systems, is very small.

In the case of most interest, which occurs very frequently, the transient impulse amplitude greatly exceeds the carrier amplitude. If the desired carrier has a frequency equal to the center frequency of the selector, then the transient and the carrier have a fixed phase during any one pulse. This results in a pulse of phase modulation which can have a maximum phase displacement of approximately 180° when the transient and the desired carrier are almost out of phase.

If the desired carrier has a frequency different from the center frequency of the selector, then the transient and the carrier will slip in phase between the beginning and end of any one transient pulse. For certain conditions of starting phase, this case will still result in a pulse of phase modulation as shown in Fig. 8. However, there are certain conditions of starting phase such that the resulting signal vector snaps back to its original phase after going through 360° of phase displacement during the pulse (see Fig. 8). This produces a step of phase modulation instead of a pulse of phase modulation. Thus, when the desired signal is not exactly on tune either a pulse or a step of phase modulation can result from a strong impulse noise, with the probable occurrence of the phase step becoming greater as the signal is further detuned.

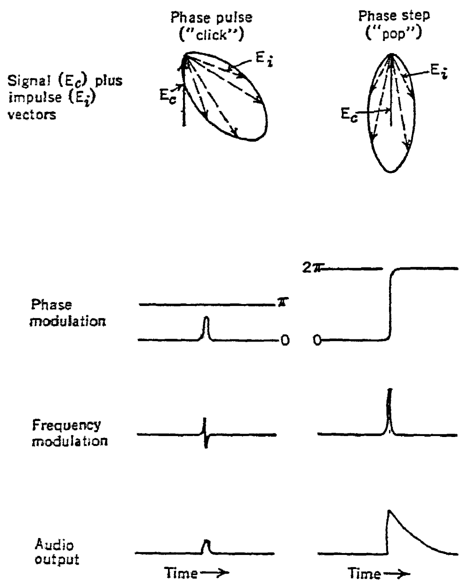


FIG. 8. Impulse Noise Interference

When a phase pulse is applied to an ideal f-m detector, the output signal is a double-polarity pulse. This double-polarity pulse applied to the de-emphasis filter and audio system results in a unipolarity pulse having relatively little energy and a short duration determined by the cut-off frequency of the audio system. This weak audio output noise is sometimes called a "click." When a phase step is applied to an ideal f-m detector a unipolarity pulse results. This pulse applied through the de-emphasis filter and audio system gives a pulse with an exponential decay determined by the de-emphasis time constant and thus having relatively more audio energy. This louder audio noise is sometimes called a "pop." *

Thus, when a strong impulse is applied to an f-m receiver, either a noticeable pop or weak click may result in the audio output, with the probable occurrence of the "pop" being directly related to the detuning of the desired carrier relative to the center frequency of the selector. In broadcast f-m with 75-kc deviation and 75-microsecond de-emphasis time constant, the click may have a peak amplitude between zero and about 6 per cent of full modulation, and the "pop" will have a peak amplitude of about 18 per cent of full modulation. The amplitude and probability of occurrence of the pop is almost independent of the amplitude of the impulse after it exceeds the carrier level by several times.

To obtain the performance described above, the f-m receiver must have good a-m rejection; otherwise the large amplitude modulation resulting from the impulse noise will be heard. Also, care must be taken to see that the receiver recovers immediately after a strong impulse; otherwise the absence of a signal immediately after an impulse may result in a large audio output due to inadequate downward a-m rejection. This trouble can result particularly in a grid-bias limiter with an improper grid time constant. Another limitation preventing ideal performance can be spurious phase modulation produced within the receiver, during the impulse, from such causes as change in input capacity of amplifier tubes.

* The Theory of Impulse Noise in Ideal Frequency-modulation Receivers, D. B. Smith and W. E. Bradley, *Proc. I.R.E.*, Vol. 34, 743 (October 1946).

SECTION 9

PULSE TECHNIQUES

PULSES AND PULSE SYSTEMS

By Harold A. Wheeler

PULSE CIRCUITS

By J. J. Okrent

By Harold A. Wheeler

PULSE TECHNIQUES

PULSES AND PULSE SYSTEMS

By Harold A. Wheeler

1. INTRODUCTION

The various uses of pulses in signaling and measurements were greatly advanced during World War II with the advent of numerous devices for aiding navigation, detecting and locating enemy craft, and performing difficult measurements and computations. It is the purpose of this section to give a broad perspective on the many applications of electronic pulse techniques and their limitations, together with a few of the more common circuits and a large bibliography for further reference.

Pulse coding is exemplified by the time-honored telegraph codes, which were originally operated slowly enough for crude mechanical devices, manual transmission, and auditory reception. Electronic pulse techniques were adapted to code systems for amplifying weak signals and expediting the various processes.

Pulsed radio waves date back to the original spark transmitters of Hertz and others, which set the pattern of early radio communication. With the obsolescence of spark transmitters began the evolution of electronic pulse transmitters, which had their greatest use in "radar" during the war. They now develop as much as a megawatt of pulse power at frequencies around 3000 megacycles (wavelength 10 cm). In some cases, the old rotary spark gap has been revived to key the new magnetron pulse transmitters.

2. COMPARISON OF CONTINUOUS WAVES AND PULSED WAVES

Various kinds of information, such as voice or music, are transmitted by corresponding modulation of a carrier wave. (See Section 5, Transients in Networks, and Section 17, Telephone Systems.) In the simplest form, the carrier is a continuous wave of a fixed frequency, and its amplitude is modulated in accordance with the sound wave or other information to be transmitted. Amplitude modulation is unique in that the modulated wave can be transmitted within the narrowest bandwidth in the frequency spectrum. Other forms such as frequency modulation and pulse modulation require excess bandwidth but in return they secure some advantages which may justify the cost in bandwidth.

The continuous waves used in amplitude or frequency modulation on one hand, and pulsed waves on the other hand, have entirely different properties which require different points of view in their application. These differences are most pronounced in the selection of one signal out of several signals or noise of comparable amplitude.

Figure 1 shows the principles of selection between two signals, regardless of their relative amplitude. As modulated continuous-wave signals are coextensive in time, frequency selection must be used for each channel by means of band-pass filters. Pulsed signals, however, can be separated in time, and so it is possible to use time selection as well as frequency selection for filtering one channel from another. The "skirt selectivity" in frequency selection denotes the attenuation just outside the desired frequency band; in time selection it denotes the rate of damping of one pulse to clear the way for the next pulse of another channel. Frequency selection is subject to harmonic interference, as shown in dotted lines in Fig. 1; the analogous interference in time selection is caused by pulse echoes in the transmission paths in enclosed circuits or open space.

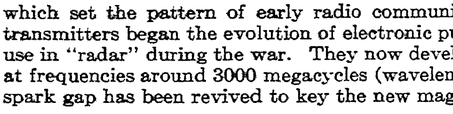


FIG. 1. Selection in Frequency or Time

In any system including several signals with the same form of modulation, some severe requirements have to be met in order to avoid interference between signals. Continuous-wave signals require that the response be very nearly linear in order to avoid harmonic interference or cross-modulation of one signal by another. This is not a requirement for pulses separated in time, because they are not coexistent. Instead, the pulse signals require that each pulse be damped out immediately after its occurrence and that later echoes be avoided.

The rating of equipment for pulse modulation places the emphasis on peak values rather than average values. For example, small vacuum tubes can be made to tolerate high peak values of current and voltage if they occur during only a small fraction of the time. Accumulative effects, such as heating and the decomposition of the glass, become less important because they depend on average values.

3. TYPES OF PULSE MODULATION

The many possible ways of modulating pulses involve three basic types of modulation as illustrated in Fig. 2. Height modulation (*a*) corresponds to amplitude modulation of a continuous wave. Width modulation (*b*) and spacing modulation (*c*) involve only the time dimensions and are therefore not critically dependent on the pulse amplitude.

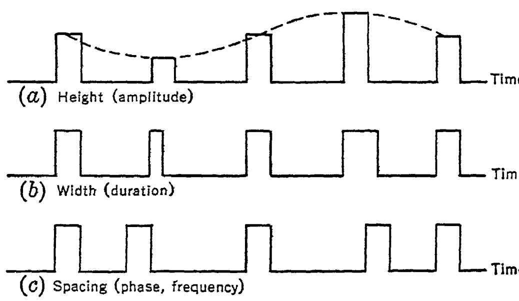


FIG. 2. The Three Basic Types of Pulse Modulation

The greatest advantages of pulses are realized in time modulation (*b*) or (*c*) as distinguished from amplitude modulation (*a*) in Fig. 2. The telegraph codes are an example of width-and-spacing modulation. It is permissible to use amplitude clipping or limiting circuits, since the amplitude need not be preserved. Also the detectors are made responsive to timing and can be made unresponsive to amplitude fluctuations such as power-supply ripple.

Pulse echo systems, such as radar, utilize the timing of the echo to determine the distance. Some pulse systems use directive antennas which receive alternately on two crossed lobes of the directive pattern. In this case, the relative amplitude of echo pulses must be preserved and the direction of reception is observed at the intersection of the two lobes by equalizing the echo-pulse amplitudes.

Some kinds of information, such as numbers, can be transmitted by grouping together a number of pulses in succession. Each group can be evaluated by a pulse counter. Multiple-pulse coding is essentially similar to pulse-width modulation but has some advantages in handling and in reliability of decoding.

The modulation of pulses of uniform width is similar to the modulation of a subcarrier, which in turn modulates a carrier. The pulse frequency is intermediate between the modulation frequencies and the carrier frequency, as is a subcarrier frequency. In the case of pulses, however, several sets of short pulses of the same frequency can be superimposed for multiplexing simply by displacement in time, whereas each continuous-wave subcarrier would need a different frequency. The pulse pattern can be subjected to any method of modulation applicable to a carrier or subcarrier, notably amplitude modulation as in Fig. 2(*a*) and phase or frequency modulation as in Fig. 2(*c*).

4. SPEED OF INFORMATION

A time variation of a quantity (such as current or voltage) may be regarded as comprising a succession of contiguous pulses of varying amplitude (see Fig. 4, p. 5-28). The speed of information that can be transmitted through a signal channel by such a variation may be conceived as the maximum frequency of such pulses whose presence or absence can be individually detected. (A space is regarded as an absent pulse, or one of zero amplitude.) Therefore the speed of information is limited by the frequency bandwidth. In the case of a low-pass channel (or one-half of a double-sideband band-pass channel), the nominal minimum bandwidth is one-half the maximum pulse frequency as here conceived, but somewhat greater bandwidth is needed for insurance of pulse damping and for skirt selectivity against adjacent frequency channels. Figure 3 shows the nominal bandwidths required for a speed of information equal to $2f_c$. (The maximum pulse frequency as here used is twice the maximum frequency of discrete pulses separated by intervening spaces of equal width; for those separated pulses the nominal minimum bandwidth becomes equal to the pulse frequency.)

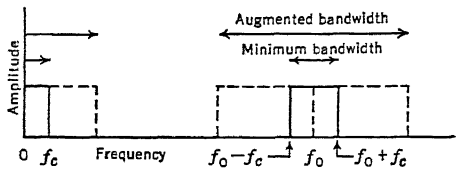


FIG. 3. Frequency Bandwidths Required for a Certain Speed of Information

The foregoing relation is based on the assumption of adjacent non-overlapping pulses, although some overlap is permissible in practice. The particular needs of the system determine how much the actual bandwidth must exceed the nominal minimum bandwidth. Figure 4 shows the distortion of a discrete square pulse which is caused by reduction of system bandwidth.

In Fig. 4 the square pulse (*a*) represents the voltage pulse or the current pulse produced by a system with a very wide bandwidth. In the remaining cases the nominal bandwidth of the system (f_c) is successively reduced to show its effect on the output pulse caused by a square input pulse. Case (*b*) shows a system bandwidth (f_c) approximately four times that of the nominal pulse bandwidth. This limitation causes sloping sides but retains a flat top over part of the pulse width. In (*c*), the system bandwidth is reduced to twice the nominal pulse bandwidth, just leaving a peak at the original amplitude. A system bandwidth equal to the nominal bandwidth of the square pulse (*d*) leaves the pulse slightly reduced in amplitude and considerably widened. Further, halving of the bandwidth (*e*) reduces the amplitude of the output pulse to less than one-half of the value in (*a*) and increases the width to more than double. Cases (*c*) or (*d*) may be regarded as practical compromises.

Like frequency modulation or subcarrier modulation, pulse modulation unavoidably increases the bandwidth requirements for the same speed of information, in the manner of the augmented bandwidths indicated by dotted lines in Fig. 3. The greater bandwidth inherently increases the average power of background noise caused by thermal agitation of electrons. It also makes possible a proportional increase in the peak power by pulsing, while maintaining the same average power.

If the signal amplitude is comparable with the noise amplitude, a change to pulsing with its greater bandwidth is no advantage. If the signal is somewhat stronger than the noise, however, and if the pulses are modulated in time, it is found possible to secure an advantage in signal-to-noise ratio which is comparable with that obtained in wide-band frequency modulation over the same bandwidth. Therefore, if the augmented bandwidth is available, pulse modulation is another way to take advantage of it.

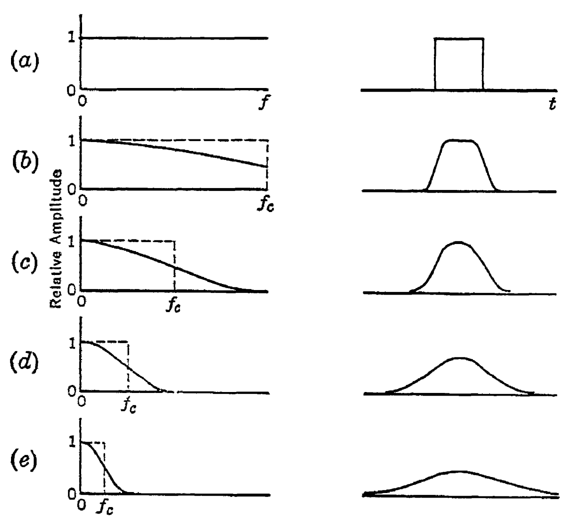


FIG. 4. The Widening of a Square Pulse by Reduction of Frequency Bandwidth

It may happen that, for some reason, more bandwidth is available than the minimum needed for the desired speed of information. At very high frequencies, the accidental frequency fluctuations of the signal may require an augmented bandwidth in the receiver. Some of this excess bandwidth may then be utilized to advantage by pulse modulation.

5. COMMUNICATION

The various types of pulse modulation have long been used in low-speed and high-speed code transmission of word messages (see Sections 17 and 18, Telephony and Telegraphy), but the greatest advance in pulse techniques has been utilized more recently in the multiplex transmission of several voice channels on a single microwave beam as a carrier. This system is taken as an example of the communication possibilities with pulse modulation.

The carrier is modulated in short pulses, and the pulse spacing is modulated by sound waves. This is the type of modulation shown in Fig. 2(*c*) above. The multiplex operation is accomplished by interspersed pulses as shown in Fig. 5. A single group of pulses comprises a sequence including one pulse assigned to each channel. One channel is reserved as synchronizing pulses to initiate each counting sequence in reception. Each of the remaining channels is modulated by shifting its pulses in time in accordance with the sound wave to be transmitted. The amount of time modulation of each pulse is limited so that the modulation of one pulse will never encroach on the time allotted to the modulation of the adjacent pulses.

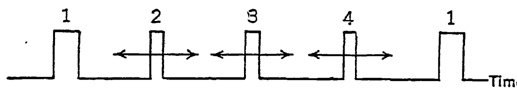


FIG. 5. Multiplex Operation of Several Channels by Pulse-time Modulation

In the transmitter, the pulses belonging to each channel are synchronized by the first channel but are otherwise separately generated and modulated. Then all channels are combined with pulses interspersed, and the composite pulse pattern is used to modulate the carrier wave.

In the receiver, the modulated carrier wave is amplified and detected, then each sequence of pulses is distributed among the several channels under guidance of the synchronizing pulses of the first channel. The distribution of each sequence may be accomplished by some form of counting or time selection. As long as the successive pulses are separated in time, there is no interference between channels. The time selection of multiplex channels offers some advantages over frequency selection, unless there are strong echoes with enough delay to overlap succeeding pulses, a condition that can be avoided by highly directive beam transmission.

Reliable reception is generally possible if the desired pulse peaks are received somewhat stronger than the peaks of noise or other interference. By amplitude limiting and clipping, the pulse peaks are flattened and the lower parts of the pulses (in the noise background) are discarded. The result is a succession of square pulses with the same timing as the edges of the received pulses. These reconditioned pulses are distributed to the separate channels for recovery of the modulation.

Since the sides of each pulse are sloping, the timing of the reconditioned pulse is still subject to some disturbance by background noise, as illustrated in Fig. 6. The noise causes some vertical displacement of all parts of the pulse, while there is no change in the level which determines the reconditioning and subsequent detection. Therefore the vertical displacement is translated to a minor amount of time displacement, always less than the pulse width. The time modulation caused by noise may be compared to the available time modulation by the signal, to determine its disturbing effect. Increasing the frequency bandwidth proportionately increases the slopes and thereby decreases the response to background noise (as in wide-band frequency modulation). Increasing the available time width for modulation (as by decreasing the number of channels) decreases the ratio of the noise modulation to the signal modulation.

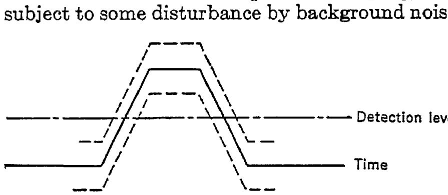


FIG. 6. Pulse Detection with Background Noise

The ultimate effect of the noise on the pulse slopes in Fig. 6 depends on the kind of detection. The simple detectors of time modulation operate on one edge of every pulse, either the leading or the trailing edge. Such detection retains the full effect of the background noise on the sloping edge of each pulse. The time detection may be designed to operate on the center of the reconditioned pulse, in which event there is approximate cancellation of those noise components that merely shift the pulse up and down, as illustrated in Fig. 6, leaving only the effect of those components that distort the shape of the pulse. The choice of the kind of detection depends not only on the noise but also on other factors which may be more important.

A common form of interference in pulse transmission is echoes caused by reflection of waves from objects in space or from irregularities in transmission lines. In communication between aircraft, the principal cause of echoes is ground reflection. Figure 7 shows how an echo may distort the trailing edge of a pulse. A direct pulse is interfered with by a slightly later echo pulse (shown in dotted lines) which is considerably weaker (below the detection level). The diagram shows only the envelope of the pulsed wave. As the relative phase of the carrier wave of the two pulses may have any angle, the echo may add or subtract on the trailing edge, as shown. Between fixed transmitter and receiver, the effect of the echo is fairly steady, varying slowly with frequency drift and environment; therefore it contributes little or no disturbance in the receiver. If the distance is variable, as between aircraft. the relative phase of the two pulses varies at random, and so an echo causes noise if the time detection operates on the trailing edge.

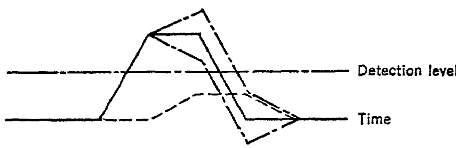


FIG. 7. Pulse Detection with Echo

Pulsed waves are most commonly obtained by pulse modulation of a carrier-frequency oscillator which delivers the required power directly to the antenna. At the beginning of each pulse, the oscillation has to build up from the noise level. Therefore the oscillator acts as a superregenerative amplifier of the background noise. The resulting noise on the leading edge of the pulse is illustrated in Fig. 8. The modulator pulse is shown in dotted lines. At the beginning of the modulator pulse, the oscillation starts to build up exponentially from the noise level and soon reaches equilibrium at the power level of the oscillator. However, the fluctuation of the noise causes a variable delay in the build-up of successive pulses, which appears as a "jitter" in the leading edge. If the time detection operates on the leading edge, the result is noise in the receiver. This effect is absent on the trailing edge because the latter is determined by exponential damping from the stable level of the oscillator on the peak of the pulse. (The noise on the leading edge can be avoided if the pulse modulation is applied to an amplifier following a continuous carrier-frequency oscillator, but this method has other disadvantages.)

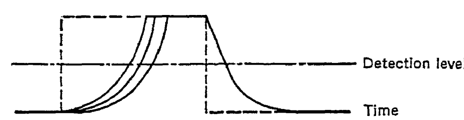


FIG. 8. Output Envelope of Pulsed Oscillator

The use of a very short pulse, shifted in time with modulation, appears to be the most economical of power while realizing the advantages of pulse modulation. The pulse duration should be nearly the least that can be transmitted within the available bandwidth in the frequency spectrum. Then the available average power can be utilized to secure greatest pulse amplitude for overcoming noise. Since the detection operates on the edges of the pulses, greater pulse duration is no advantage and greater amplitude is a proportionate advantage.

Detection on the leading edge, the pulse center, or the trailing edge, is a choice that depends on the nature of the system. In beam transmission along a fixed path, echoes are unlikely, so detection on the trailing edge is preferable to avoid the oscillator noise on the leading edge. Broadcast transmission, especially between moving stations, is subject to echo interference, which gives the advantage to detection on the leading edge; the oscillator is then designed to minimize the superregenerative noise. If both the echoes and the superregenerative noise are less than the random background noise, center detection would give the best performance and its complication might be justified.

The potentialities of multiplex pulse transmission are indicated by the studies which show that the entire broadcast services for a large city could be transmitted from a single microwave system centrally located on the highest building, with a service area limited by the optical horizon.

6. PICTURE TRANSMISSION

In picture transmission by scanning methods (see Section 19, Facsimile, and Section 20, Television), pulses are relied on not only for reproducing the picture elements but also for timing and synchronizing the scanning process. Since the common systems for picture transmission utilize scanning methods, many examples of pulse circuits and their applications are found in both facsimile and television.

Figures 9 and 10 show respectively the essential components of a picture transmitter and receiver with scanning by deflection of an electron beam, as in present-day television.

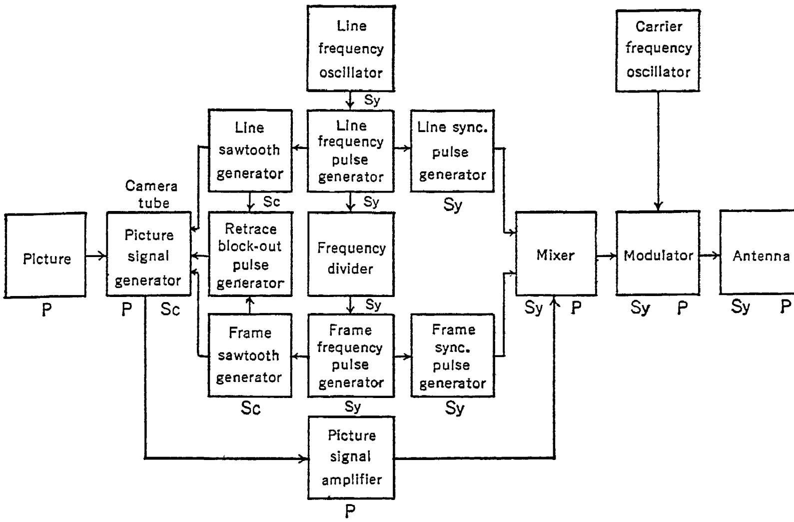


FIG. 9. Block Diagram of Picture Transmitter

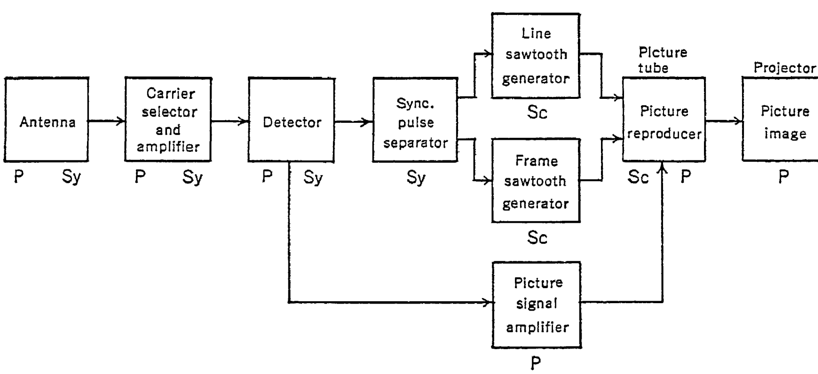


FIG. 10. Block Diagram of Picture Receiver

In these block diagrams, the various functions are coded:

P = picture channel.
Sc = scanning functions.
Sy = synchronizing functions.

The picture signal is generated in the transmitter and converted back to an image in the receiver. The scanning function is individual to transmitter or receiver. The automatic synchronizing operation is initiated in the transmitter by the line-frequency oscillator and used directly to time the scanning in the camera tube; it is maintained by transmitting timing pulses along with the picture signal, which are selected in the receiver for holding in step the scanning in the picture tube.

Figure 11 shows an example of the pulses involved in the scanning of a single horizontal line in a picture. This line is located at the dotted line in the pattern (a). The graph (b) shows in the period P the picture signal for the line. As the scanning line crosses the circular line, a black pulse is generated; as it crosses each edge of the black disk a step is generated, two steps making a wide pulse. The line signal is preceded and followed by a synchronizing pulse *Sy* which times the successive lines. The sync pulse is communicated at an "infra-black" (blacker than black) level and so it does not appear in the retrace lines back across the picture between lines. The graph (*c*) shows the sawtooth wave of voltage or current which is used to deflect the cathode ray or electron beam, in the camera tube or picture tube, from one end of the line to the other.

The transmission of the picture involves the communication of many pulses, each having an amplitude proportional to the brightness of a small element of the picture. The "speed of information" may be expressed as the number of "independent" picture elements that can be transmitted per second, or the number of pulses per second. It ranges from hundreds in facsimile up to millions in television. Completely independent transmission of adjacent picture elements by successive pulses is never attained because the electrical circuits and other devices cause distortion resulting in overlapping of adjacent pulses. One of the principal problems is therefore the design of the circuits to reproduce the short pulses with clean edges by minimizing amplitude and phase distortion over the requisite frequency bandwidth in the circuits. Both kinds of distortion are usually more prevalent at frequencies near the limits of the frequency band required for reproduction of the pulses and steps. The higher frequencies play the major part in reproducing vertical lines or edges; the lower frequencies, in reproducing long pulses or background.

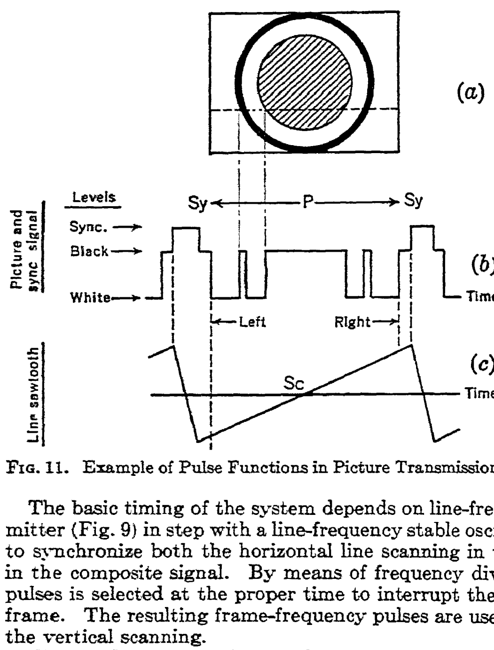


FIG. 11. Example of Pulse Functions in Picture Transmission

The basic timing of the system depends on line-frequency pulses generated in the transmitter (Fig. 9) in step with a line-frequency stable oscillator. These pulses are used directly to synchronize both the horizontal line scanning in the camera tube and the sync pulses in the composite signal. By means of frequency dividers or pulse counters, one of these pulses is selected at the proper time to interrupt the vertical scanning at the end of each frame. The resulting frame-frequency pulses are used in a similar manner to synchronize the vertical scanning.

The synchronizing pulses need to be specified and preserved in shape only to the extent required by precision of timing in the scanning operation. Since they are selected from the signal at a certain infra-black level which may fluctuate with picture content, it proves necessary to hold the edge of a sync pulse nearly as steep as the edge of a step in the picture signal. As sync pulses need not be modulated in amplitude, they can be subjected to limiting or clipping action to remove accidental changes in amplitude. A practical television system may have rather complex sync pulses for facilitating the separation of line and frame sync signals, and for minimizing their susceptibility to interference from the picture signal or other disturbances.

7. COMPUTERS

The solution of certain problems, such as the trajectory of a missile, requires elaborate calculations, which are laborious even when done on mechanical computing machines. Electronic computing machines can increase the computing speed by a factor of 1000 or more, so that a compilation of tables which would take years on a mechanical computer would be done in a few hours or days on an electronic computer. The addition of a digit in electronic computing requires a few microseconds as compared with a few milliseconds in mechanical computing.

An outstanding electronic computer is one of the products of World War II, the "electronic numerical integrator and computer" (eniac).

An electronic analog of the mechanical counter or electromechanical stepping relay may be made up of a number of flip-flop relaxation circuits connected to be switched consecutively by successive pulses. A single flip-flop circuit is analogous to a toggle switch, snapping from one condition to the other and remaining there until an external force snaps it back. In the block diagram of a two-digit ring counter, Fig. 12, each input pulse snaps the "on" switch "off" (to the left), which in turn snaps the next above switch "on" (to the right), increasing the indicated number by one. If the units column indicates nine, and a unit is added, the units "nine" switch is snapped off, the units "zero" switch is snapped on, and the tens column receives a pulse, increasing its indicated number by one. In the eniac, when going from "nine" to "zero," a "carry" switch stores the pulse for the next column. Simultaneous addition in all columns is thereby made possible. The "carry" switch is snapped off after addition in the columns is complete.

An electronic computer may consist of the following:

1. A number of counters to add and store numbers represented by pulse groups.
2. A generator of suitably timed standard pulse signals.
3. Devices for converting numbers supplied to the computer into pulse groups representing the numbers.
4. Devices for printing numbers stored in counters.
5. Devices for combining the quantities in the counters in different ways to add, multiply, divide, etc.
6. Devices to produce automatic repetition of required computations, as in numerical integration.

A basic decade ring counter was described above. The decade counter may be so arranged that, when supplied with a standard sequence of pulses from the timing generator, a number of pulses corresponding to the number stored in the counter is transmitted to another counter which adds it to its own stored number.

Various groups of pulses from the timing generator may be selected and combined by switching to produce pulse groups corresponding to numbers to be supplied to the computer. The numbers stored by a counter may be indicated by neon lamps energized by "on" flip-flops, as indicated by the black dots in Fig. 12. Card punching or printing devices may be similarly actuated to record a stored number.

Rapid switching is accomplished by means of triple-grid tubes, keying pulses being applied to one control grid and information pulses being transferred via another control grid.

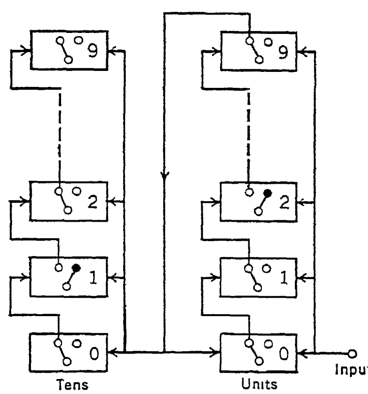


FIG. 12. Two-digit Decade Counter

The flexibility of an electronic computer as regards physical location of components, number of interconnections possible, and the large tolerances in amplitude possible, facilitates the design of computers to perform very elaborate computations with lightning speed.

Where absolute numerical precision is not required, the step counter may be replaced by simpler devices relying on continuous integration. In one type, each pulse delivers an incremental charge to a capacitor whose stored charge is indicated on a meter scale. In another type, the pulse rate is indicated by conducting the incremental charges through a d-c meter so that the current is proportional to the number of pulses per second. Examples of these types are found in Geiger counters, long in use for recording pulses of radioactive radiation.

8. DISTANCE MEASUREMENT

Distance can be measured by timing pulses transmitted through a medium in which the wave velocity is known. In the interest of precision, the duration of each pulse should be much less than the time required for the wave to cover the distance in question. The three basic methods used in operating systems are described in Section 22.

Diverse Waves. Observations of distance by this principle are based on the reception of a pulse by two kinds of waves, over the same distance, having different velocities so that the time difference in reception is a measure of the distance. Examples of light and sound waves originating in simultaneous pulses and traveling over the same distance are found in lightning and thunder or in the seeing and hearing of a distant steam whistle. In these cases, the only appreciable delay is in the sound wave, and so the computation of the distance is based on the velocity of sound in air. An example of different types of waves in the same medium is given by the seismograph of an earthquake. The pulses are transmitted through the earth by longitudinal waves (pressure waves, like sound) and transverse waves, having different known velocities so that the difference in time of reception permits computation of the distance. (See Section 22, article 9.)

Pulses from Diverse Locations. One of the outstanding radio navigational systems (loran, gee, and shoran, Section 22, article 10) also utilizes pulses spontaneously transmitted to the observer, but from widely spaced points and carefully synchronized in time. The time difference in their reception by radio waves permits precise computation of the differential distance relative to each pair of spaced points, and such observations on two different pairs determine the position of the receiver.

Reflection or Return of Pulse. The more versatile systems employ pulse waves transmitted from the observer to a distant point and back again to his receiver. The elementary sonar and radar systems utilize short pulses of sound or radio waves transmitted toward an object and reflected back to a receiver (Section 22, article 10). Special radar systems for beacons or identification rely on a pulse repeater at the object, which receives and retransmits the pulses with coding of some kind (radar beacons, Section 22, article 10; and Lanac, Navar, and Teleran, Section 22, articles 6 and 11). In any case, the round-trip time at known velocity determines the distance.

Table 1. Wave Velocity

Kind of Wave	Medium	Velocity
Electromagnetic (light, radio)	Free space (and air)	300 m/μs
Sound	Air (atmospheric pressure, 20 deg C)	344 m/sec
Sound	Water (20 deg C)	1464 m/sec
Seismic (longitudinal, sound)	Earth	4–14 km/sec
Seismic (transverse)	Earth	3–10 km/sec

Table 2. Characteristics of Radar

Carrier frequency	30–30,000 Mc
Carrier wavelength	10–0.01 meters
Pulse width	10–0.25 μs
Pulse repetition frequency	60–4,000 cps
Pulse power	10–1,000 kw

Table 1 shows various values of the wave velocity in different mediums. Table 2 shows the approximate range of pulse characteristics in radar systems of the reflection type. Since the error of time measurement may be reduced to a fraction of the pulse width, the corresponding distance error may be reduced to the order of 1–0.01 mile. The range of echo reception varies from a few miles to above 200 miles.

9. PULSE MEASUREMENTS

PULSE AMPLITUDE. A simple pulse of voltage or current can be displayed on an oscilloscope and its various properties determined from the calibration of the scope. Such a display is shown in Fig. 13. The small pulses or "pips" are superimposed marker pulses carefully timed to provide a time scale. A representative scale might have small marker pips every microsecond, and every fifth one enlarged. In Fig. 13 the pulse under observation is 1 division wide at the peak and about 2 divisions at the base; it might be rated 1.5 divisions wide at half amplitude.

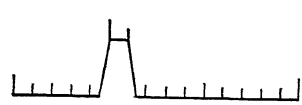


FIG. 13. Oscilloscopic Observation of Pulse Characteristics

A pulsed wave having many cycles of the carrier in a single pulse presents a special problem if the scope cannot give a calibrated display of the actual carrier cycles. The wave must be rectified for display, and the performance of the rectifier is difficult to predict or measure. If the rectifier responds quickly, the rectified pulse may be displayed as in Fig. 13, which is adequate for observing its width or duration.

An indirect method is usually employed for rough measurement of the amplitude of a repeating pulsed wave, the pulse power, for example. The average power is measured by a thermal device such as a thermocouple or bolometer. The pulse duration is observed as in Fig. 13. Then the ratio of peak to average power is equal to the ratio of the period of repetition to the pulse duration. This method is accurate if the time occupied by the sides of the pulse is much less than the duration of its peak. Otherwise the pulse width is indefinite, usually approximated by the width between the points at one-half the peak power or the peak amplitude. (This same method is applicable to simple pulses if there is no background of direct current or voltage, but it is not usually needed in this case.)

In the case of a pulsed carrier wave, some care is required in expressing the peak values during the pulse. The peak power is the mean power of the carrier wave at the peak of the pulse. If there are minor ripples on the flat top of a pulse, the pulse power is stated at the level of the flat top. The peak voltage may be stated as the peak value of the carrier voltage at the level of the peak or flat top, since that is the value significant for voltage breakdown. The pulse current in a vacuum tube carrying a pulsed wave is usually stated as the average value of the current during the peak or flat top of the pulse, since that is the value readily measured by means of an oscilloscope.

The pulse amplitude of a pulsed wave is best measured by comparison with a continous wave of known amplitude, since such a wave presents no unusual problem. The "notch" method shown in Fig. 14 is based on this principle. The comparison wave is cut off or notched for the duration of the unknown pulse. Then the pulsed wave is superimposed thereon, and the composite wave is rectified and displayed on the scope. The two amplitudes are equalized to give the appearance of Fig. 14; then the pulse amplitude is known. This method is independent of the rectifier and scope characteristics and has been found very useful. Measurements have been made as low as 200 μw peak pulse power and 2 μw average power.

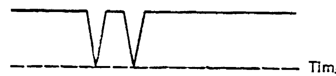


FIG. 14. Measurement of Pulse Amplitude by Comparison with Continuous Wave

PULSE DURATION. The simplest method of observing pulse width or duration is that shown in Fig. 13.

Another method is based on the frequency spectrum of a pulse, illustrated in Fig. 15. (The method of observing the spectrum is to be described below.) The frequency spectrum of a pulse has a width inversely proportional to the width or duration of the pulse. Furthermore, if the pulse has steep sides, the spectrum has a sharply defined minimum value at a frequency differing from the maximum by $2f_c$, as shown in Fig. 15 for a pulse or a pulsed wave. The frequencies of minimum response can easily be observed by a sharply tuned receiver. f_c having been determined, the pulse duration is $1/2f_c$. For example, f_c is $^1/_2$ megacycle for a pulse width of 1 microsecond, and the minimum response is displaced 1 megacycle from the maximum response.

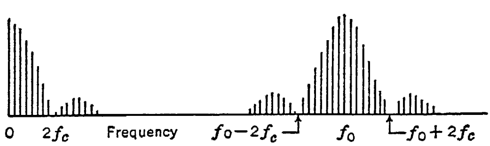


FIG. 15. Frequency Spectrum of Pulse

If it is convenient to measure the peak and average values of the power or voltage or current of a repeating pulse, the duration may be computed. The ratio of the pulse duration to the period of repetition is equal to the ratio of average to peak values, commonly called the pulse "duty cycle."

AVERAGE VALUES. A repeating pulse has a definite average value of power or voltage or current. Any one of these average values may be significant in determining the heating or other accumulative phenomena in a circuit or vacuum tube. For voltage or current, it is important to specify whether the average or root-mean-square value is to be measured.

The average power, or the rms value of voltage or current, can be measured by a thermal instrument with a time constant much greater than the period of repetition of the pulses. The peak voltage during the pulse is abnormally large for the usual instrument of this type, and so voltage breakdown may occur, in which event this defect may be corrected by redesign.

The thermal instruments in common use include the thermocouple, the bolometer bridge, the lamp with photometer, and the resistor with calorimeter. The most severe requirements are met in measuring pulsed waves of ultra-high carrier frequencies, say 1000 megacycles and upward.

The bolometer bridge is a d-c four-arm bridge with a temperature-sensitive resistor in one arm which receives the pulse power to be measured. This device has been improved by the use of a composition resistor very sensitive to temperature, called the "thermistor." Since it is desirable for the measuring circuit to present constant resistance, the bridge is rebalanced by decreasing the d-c level in all arms to restore the variable resistor to the same temperature, establishing a d-c calibration of the power sensitivity in normal operation. The device is sensitive to low power of the order of 1 milliwatt and also can be adapted to greater power.

The lamp is useful for medium power of the order of 1 watt. The radiation from the lamp is indicated by a nearby photocell and meter. It has two great advantages: it requires no electrical connections between the pulse circuit and the indicating circuit, and it can be calibrated by direct current. Its one great disadvantage is its large variation of resistance at the high temperatures incidental to radiation of light.

The calorimeter is useful for high power of the order of 1 kilowatt. It can be designed for nearly constant resistance at very high frequencies, since resistance variation is not essential in its operation at moderate temperatures.

A possible alternative to the thermal meter is a square-law rectifier, which may be approximated by proper design of a vacuum-tube rectifier. It is important to hold the peak value within the range of square-law operation, which severely limits the utility of this type of instrument.

The average value of current or voltage of repeating pulses, assuming zero between pulses, can be measured in an ordinary magnetic d-c meter having a time constant much greater than the period of repetition. In a voltmeter, care must be taken that the pulse voltage is not excessive.

PULSE FREQUENCY. Repeating pulses commonly have a steady value of the pulse repetition frequency (prf). However, a receiver or a replying transmitter may have a random pattern of repetition, depending on the traffic, and then it may be desirable to have a continuous indication of the average pulse frequency over a short period such as 1 sec in order to be aware of overloading.

A constant pulse frequency is easily observed. It is usually an audible frequency and can be compared with a calibrated audio frequency. The integrating type of frequency meter which is used for direct-reading audio-frequency meters may be designed for pulse-frequency measurements. It gives the continuous indication desirable for monitoring a varying pulse frequency. If the pulse amplitude and width are uniform, the same result can be obtained by inserting a d-c or average-power meter in the circuit where it will indicate an average value proportional to the pulse frequency.

PULSE DETAILS. A critical analysis of the details of a repeating pulse or group of associated pulses may require a display on the oscilloscope with a greatly expanded time scale. For example, the entire width of the scale may be only a few microseconds, even though the pulse groups may be separated by a millisecond.

The synchroscope is a special oscilloscope designed for this purpose. Each trace or sweep is triggered by the first edge of the pulse pattern to be observed, and so close registry of successive traces is assured, hence the name "synchroscope." At the end of each trace, the spot waits for the signal to start the next trace. Care is taken to insure that each trace starts at the same point and proceeds at the same rate. Any failure of registry then indicates variations in the pattern.

FREQUENCY SPECTRUM. A simple pulse or a pulsed wave has a frequency spectrum as illustrated in Fig. 15, the former centered on zero frequency and the latter on the carrier frequency. The significance of the spectrum is the ability to excite a circuit which selects a bandwidth much less than the width of the spectrum.

The spectrum analyzer is a device for displaying the frequency spectrum on the scope in the form shown in Fig. 15. As the trace progresses horizontally along the frequency axis, a narrow-band receiver is tuned over the frequency range. The repeating pulse is applied to the receiver, and the rectified output is shown by vertical deflection. The pulses therefore appear as successive vertical lines whose heights show the relative amplitude of the frequency spectrum. The spectrum is a measure of the frequency bandwidth required for reproduction of the pulse, as well as the interference that may be caused in adjacent frequency channels. In the case of a pulsed wave, a symmetrical spectrum indicates pure amplitude modulation free of frequency modulation.

As appears from the width of the frequency spectrum, a pulsed wave does not give a sharp indication of resonance in a sharply tuned wavemeter. The carrier frequency is best defined by the center of a symmetrical spectrum. Therefore it is customary to provide the analyzer with a very sharply resonant calibrated circuit whose frequency can be adjusted to put a narrow gap in the center of the spectrum and thereby to determine the carrier frequency. The calibration of the trap circuit is made on the narrow spectrum of a continuous wave or long pulses of known carrier frequency.

(For oscilloscope technique, see Section 11, Wave Analysis, and Section 20, Television.)

PULSE CIRCUITS

By J. J. Okrent

10. FREQUENCY MULTIPLIERS, DIVIDERS, AND COUNTERS

The abrupt changes in amplitude associated with pulse phenomena give rise to a wide-band frequency spectrum, as indicated above (see Computers, article 7 above; also Section 11, Frequency Measurements). The frequency spectra of certain periodic wave forms, as

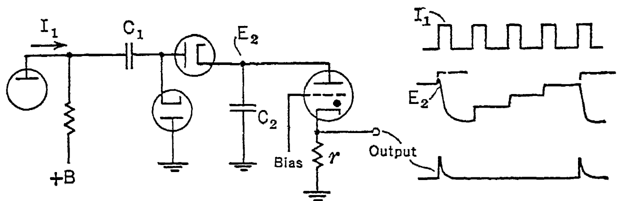


FIG. 1. Counting Type of Frequency Divider

expressed by Fourier series, are well known. Periodic pulses have strong components of high-order harmonics. These are produced for frequency comparison by making a stable sinusoidal oscillator synchronize a relaxation circuit which produces the high-order harmonics of the stable frequency. The harmonics are compared (by zero beat, for example) with a signal whose frequency is to be standardized. Substantial output of a harmonic frequency is obtained by making the relaxation pulse excite a circuit sharply resonant at a chosen harmonic frequency. Pulse circuits incidentally radiate interfering power at harmonic frequencies unless they are adequately shielded and filtered.

Frequency division may be accomplished by synchronizing a relaxation oscillator at an integral submultiple of the synchronizing signal frequency. The natural period of the relaxation oscillator is made somewhat longer than the interval occupied by the selected number of synchronizing pulses, and so the pulses expedite the relaxation in each cycle.

Another method of frequency division utilizes a "ring-counter" circuit which counts the required number of pulses and then generates a trigger pulse and starts counting over again. In Fig. 1, successive pulses of current I_1 increase the charge on the capacitor C_2 and thereby its voltage E_2, for each pulse until the output thyratron conducts, discharging C_2 and producing an output pulse.

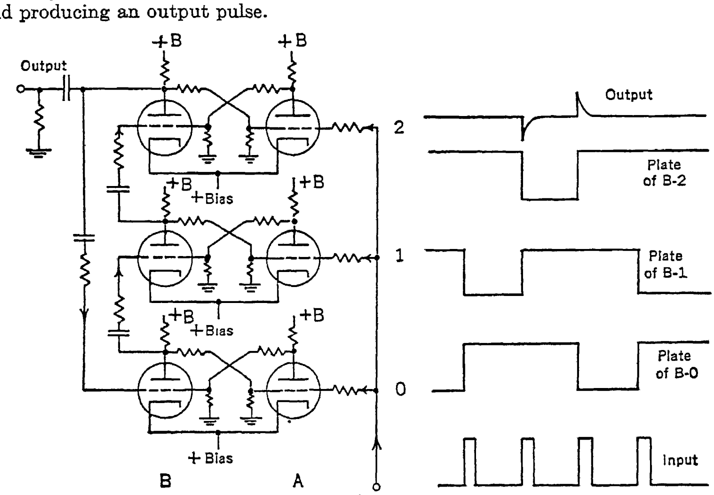


FIG. 2. Flip-flop Circuits in a Counter

Figure 2 shows flip-flop circuits of a type used in counters and computers. Successive input pulses, not necessarily at regular intervals, switch successive flip-flops to an indicating position. In each flip-flop, either triode, A or B, may conduct; the other triode is then cut off. Also, if triode A-0 is cut off, triodes A-1 and A-2 are conducting. A positive input pulse will then switch triode A-0 into conduction, and the coupling between the plate of B-0 and the grid of B-1 will switch triode A-1 into conduction. The control grid of triode A-1 is thus made sensitive to the next input pulse, and the high plate voltage on B-1 may actuate for that digit an indicating device such as a neon lamp. The coupling between triodes B-2 and B-0 completes one ring, so that the cycle is repeated every three input pulses. Output pulses with a frequency one-third that of the input pulses are thereby obtained.

11. PULSE AMPLIFIERS

The requirements to be met by pulse amplifiers are generally similar to those of television carrier-frequency and video-frequency amplifiers (see Section 7, Wide-band and I-f Amplifiers). The amplifiers must pass the essential frequency components of the signal to be amplified, with uniform gain and time delay. Any excess bandwidth passes needless background noise. A compromise choice made in radar systems allows a video bandwidth (in megacycles per second) equal to the reciprocal of the pulse duration (in microseconds), and double this bandwidth for the modulation sidebands of a pulsed carrier. This is twice the nominal minimum bandwidth defined on p. 9-11. Some additional carrier-frequency bandwidth is added for tolerance of detuning from various causes such as frequency drift.

The amplitude-and-phase spectrum concept (see Section 5, article 13) is useful in specifying low-pass and band-pass amplifiers for pulses and pulse-modulated carrier waves.

Fast recovery of normal operation after overloading by strong signals is required in systems like radar, in which a weak pulse may immediately follow a strong pulse. Figure 3 shows an i-f amplifier stage having quick recovery. The grid bias returns to normal

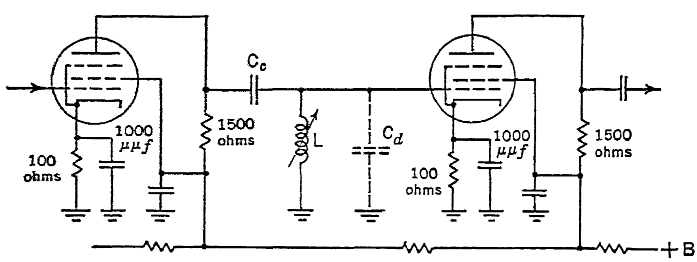


FIG. 3. Intermediate-frequency Amplifier

immediately after a strong pulse because there is negligible d-c resistance between grid and ground, and the cathode-circuit time constant is only 0.1 microsecond. The cathode-circuit time constant is made the minimum consistent with sufficient bias and bypassing. Plate current is supplied through the damping resistor for the stage. Fast recovery of normal plate voltage after a strong pulse is relatively unimportant because the operation of a pentode tube does not depend critically on plate voltage. The inductance L is made to resonate the distributed capacitance of the wiring and tubes, C_d, without any added lumped capacitance, in order to obtain maximum stage gain with the required bandwidth.

Video pulse amplifiers are frequently specified in terms of the permissible distortion of rectangular pulses by the amplifier, rather than by amplitude and phase characteristics. Thus, it may be required that a specified input pulse after amplification shall have a certain maximum rise time, fall time, ripple ratio, etc.

Resistance-coupled stages are used for voltage amplification. Because high video frequencies are usually involved, high-transconductance pentodes and relatively small coupling resistors are used. The product of the coupling resistance by the shunt capacitance places a lower limit on the time of rise and fall of output pulses. Shunt capacitance of un-bypassed components and wiring should be made as small as practicable.

In Fig. 4, an input pulse having zero rise time would produce at the following grid an amplified pulse having a leading edge rising exponentially with a time constant $\dfrac{r_1 r_2 C_d}{r_1 + r_2}$ (in ohms, microfarads, and microseconds). The amplified pulse would reach 90 per cent of peak amplitude in 2.3 times the time constant. The plateau of the pulse across r_2 decays exponentially with the time constant $(r_1 + r_2)C_c$, ordinarily much longer than T, the duration of the pulse. If, for example, less than 1 per cent decay is permissible from this coupling alone, make $(r_1 + r_2)C_c$ greater than $100T$.

The high-frequency response of resistance-coupled amplifiers may be improved by compensation and filter techniques discussed elsewhere in this book (see Section 7, Wide-band Amplifiers, and Section 20, Television).

When power amplification or impedance change is required, transformer coupling may be used. High-power pulses are economically obtained from tubes having low average-power ratings by permitting space current to flow only during pulses. The control grid

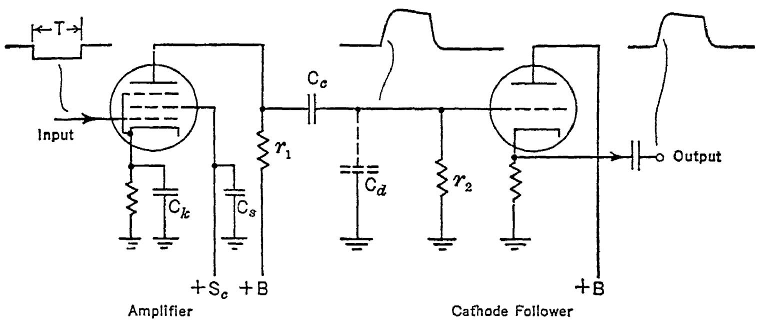


FIG. 4. Video-frequency Amplifier

in a high-power amplifier is therefore biased negative and driven positive during the pulses. Transformers may be used to provide the polarity inversion which may be required in successive high-power stages and also to permit impedance matching. Pulse transformers will be described below.

Cathode followers (see Fig. 4) are useful in obtaining high-voltage positive pulses across low impedance such as the coaxial lines commonly used. In successive stages of high-power pulse amplifiers, cathode followers make transformers unnecessary for polarity inversion. However, transformers are necessary if voltage amplification is required in addition to the current amplification obtained in a cathode-follower stage.

12. PULSE SHAPING CIRCUITS

Many of the wave forms used in pulse work are obtained by means of exponentially changing voltages and currents and by clipping or limiting action in the tubes. In sharp contrast with linear-amplifier practice, the operating conditions of the tubes in pulse shaping circuits are so chosen that grid current may flow or plate current may be cut off in order to distort a pulse or to reconstruct a clean pulse.

Clipping or Squaring. In Fig. 5, a sinusoidal voltage is applied to the grid of a tube through a high resistance. As the grid goes positive, the grid-to-cathode resistance falls

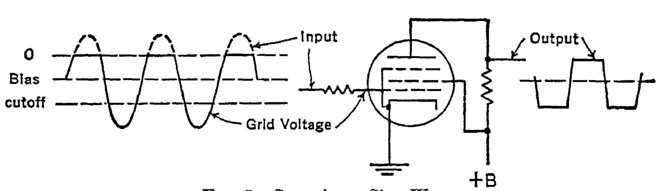


FIG. 5. Squaring a Sine Wave

abruptly, flattening the positive peak of grid voltage. As the grid goes negative, the plate current is cut off, flattening the negative peak. The output wave form is approximately square and can be squared further by added stages.

Pulse Narrowing. In Fig. 6a the time constant r_1C_1 is short as compared with the duration but not as compared with the time of rise and fall of the applied pulse, e_1. The voltage across r_1, from the step at each edge of the pulse, decays exponentially. The pulse is said to be differentiated, because the voltage across r_1 is approximately proportional to the derivative of the applied pulse voltage. Either the leading or the trailing edge may be used by the following stage. If the bias on the tube T_2 is large enough to cut off the plate current, only the positive (leading) impulse is amplified. If the bias is zero and r_2 is large, only the negative (trailing) impulse is amplified.

Long pulses may be shortened to a predetermined duration by application through a high resistance (in Fig. 6b the plate resistance of the pentode) to a delay line which is in

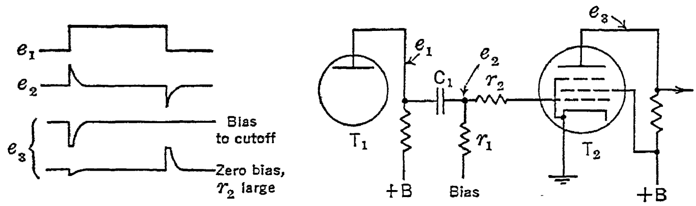


FIG. 6a. Narrowing a Pulse by *R-c* Differentiation

parallel with a resistance equal to the characteristic impedance z_0 of the line. The far end of the line is short-circuited, so that the pulse at the input end of the line is canceled by a reflected pulse of opposite polarity in twice the one-way delay time, t_d, of the line. The undesired pulse of reverse polarity which occurs at the end of the input pulse is clipped in a succeeding circuit.

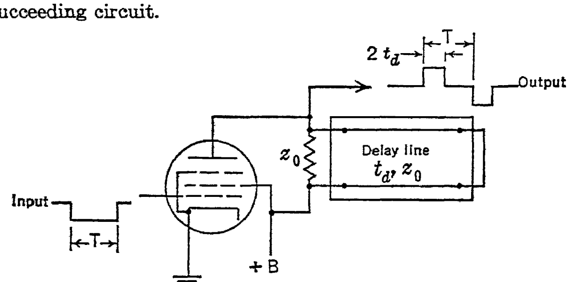


FIG. 6b. Narrowing a Pulse by Use of a Delay Line

Another arrangement for obtaining a short pulse of predetermined duration from a long pulse is shown in Fig. 6c. A half sine wave of short duration is produced in the plate circuit by the application of a long pulse. The period of the half sine wave is determined by the inductance of the coil and the distributed capacitance across it. The diode damps the oscillation on the second half cycle, reducing the oscillation after the first half cycle to a negligible level.

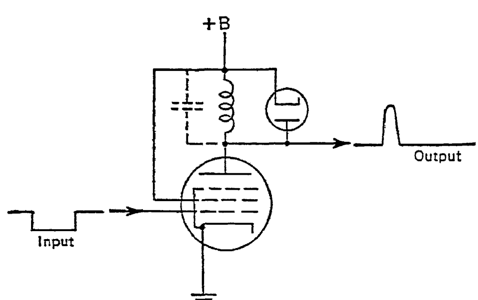


FIG. 6c. Narrowing a Pulse by Use of an Oscillatory Circuit

Pulse Widening. Integration is said to occur in Fig. 7. A positive pulse on the grid of tube *A* quickly discharges capacitor C_1, which recharges exponentially with the time constant r_1C_1. Resistance r_2 is assumed large as compared with r_1, and r_2C_2 large as compared with r_1C_1. Tube *B* is cut off during enough of the exponential discharge to provide the required pulse duration. The wave forms in solid line are typical of the circuit shown. By connecting r_2 to a positive voltage instead of to ground, a shorter duration and time of fall is obtained, as shown by the wave forms in broken line.

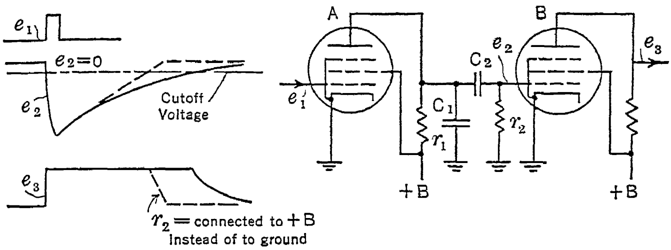


FIG. 7. Widening a Pulse

Clamping or D-c Reinsertion. When it is required that the baseline of a pulse wave form remain at a fixed voltage in spite of capacitive coupling and changing wave form, the arrangement shown in Fig. 8 may be used. If the two wave forms shown are coupled by capacitor C and resistor r_1 without the diode, the voltage limits vary as shown because the average voltage must remain zero. The use of a diode results in a short coupling time constant for negative voltages, so that the wave extends almost wholly in the positive direction. The resistance of r_1 is much greater than the resistance of the diode in the conducting direction (r_d). If no grid current flows, the rectifying action of the diode maintains on the capacitor a sufficient charge to hold substantially the entire wave positive

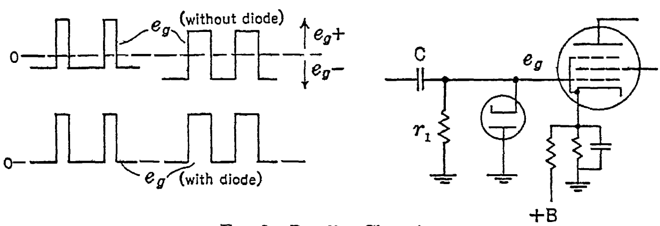


FIG. 8. Baseline Clamping

relative to ground. The diode also permits fast recovery of normal bias if grid current is drawn by a large positive pulse, because a negative voltage is quickly discharged through the diode.

13. RELAXATION CIRCUITS

Relaxation circuits are oscillators in which little if any energy is stored from one relaxation cycle to the next. The circuits have two conditions in which they are at least temporarily stable. When one condition becomes unstable, the oscillator shifts abruptly to the other condition.

An example of a circuit which has two permanently stable conditions is shown in Fig. 9. One tube is conducting and the other cut off in each of the stable conditions. Conduction

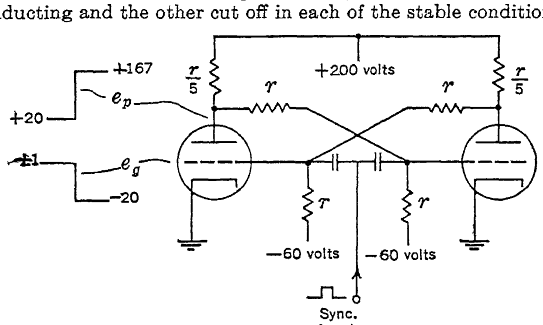


FIG. 9. Flip-flop Circuit Stable in Either Condition

is switched back and forth by successive trigger pulses, hence the designation "flip-flop" circuit. Typical voltages are shown in the figure.

In Fig. 10, a capacitor replaces one of the plate-to-grid coupling resistors, and direct coupling is used between the other plate and grid. Plate current cut off in tube T-2 and conduction in tube T-3 make a permanently stable condition. A trigger pulse through tube T-1 makes tube T-2 conduct and tube T-3 be cut off in a temporarily stable condition

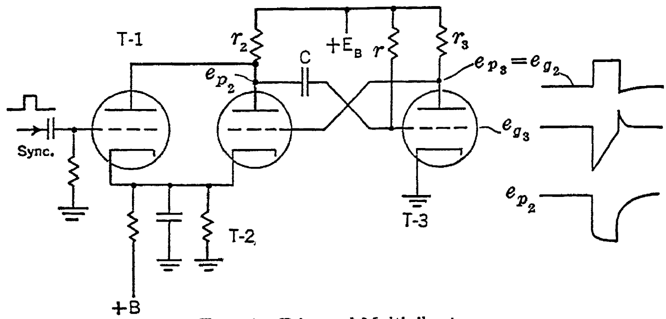


FIG. 10. Triggered Multivibrator

which ends when C discharges sufficiently through r. The circuit then switches back to the permanently stable condition. The circuit is known as a triggered or one-pulse multivibrator, or a "univibrator."

The circuit shown in Fig. 11 has only temporarily stable conditions and therefore runs free. It may be synchronized at the trigger frequency or integral submultiples thereof.

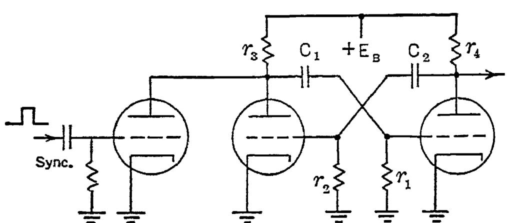


FIG. 11. Free-running Multivibrator

A transformer is used (instead of a second tube) for feedback polarity reversal in the one-tube blocking oscillator circuit shown in Fig. 12. It is free running unless the C bias is increased to limit the oscillation to one pulse. A temporarily or permanently stable condition exists between pulses while the grid capacitor is charged negative beyond cutoff. The other extreme condition exists during the pulse while the rate of increasing

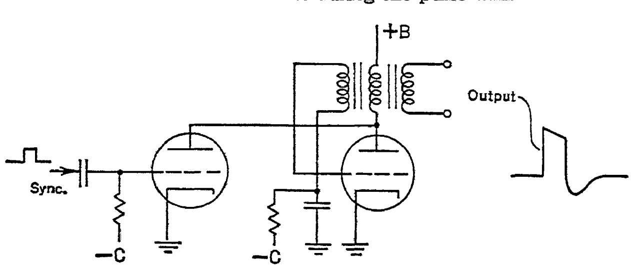


FIG. 12. Blocking Oscillator

plate current is limited by the inductance in the plate circuit; this condition causes grid current to charge the grid capacitor ready for the blocking period. Because plate current flows only during the pulses, the blocking oscillator works economically with the low impedances and high currents necessary for short pulses across the inherent capacitance.

Ionizing discharge tubes, such as thyratrons and gas diodes, are used in generating trigger pulses and sweep voltages by periodic discharge of a capacitor each time the capacitor is charged to the ionizing potential of the discharge tube. The capacitance and the charging impedance are chosen to give the desired period of pulse repetition.

A number of relaxation circuits particularly adapted for different applications are shown in other parts of this section.

14. PULSE TIMING CIRCUITS

Various timing problems arise in pulse systems. Pulse repetition rate, duration, and delay relative to a reference pulse require measurement and control.

Voltage wave forms frequently used for pulse timing are the sinusoidal, the exponential, and the rectangular. The three wave forms may be compared for basic stability resulting from their use.

With a sinusoid, if the timing action occurs as the sinusoid passes through the value zero, the timing is not greatly affected by the amplitude of the sinusoid, and stability of the same order as that of the inductance and capacitance is obtained.

In commonly used exponential timing circuits, the timing action occurs as the variable (voltage or current) reaches a predetermined amplitude. The timing is therefore a function of the initial amplitude as well as the time constant of the exponential change. Stability of the same order as that of the passive circuit elements requires compensation for changes in tube characteristics and operating voltages. Stability is generally improved by using only the early and most rapidly changing part of the exponential wave. Timing in relaxation oscillators is generally done by exponential changes to predetermined amplitudes.

A rectangular wave sent through a delay line yields a timing wave in which the entire amplitude change occurs at the most useful time. The stability obtained is limited by the allowable volume and complexity of the line.

Repetition Rate. A conventional sine-wave oscillator followed by a relaxation or shaping circuit may be used to obtain synchronizing pulses having a stable repetition rate and short rise time. The sine wave may be converted to a square wave by several clipper stages as discussed in article 12. The square wave may then be differentiated to obtain a suitable trigger pulse.

Having a sinusoidal oscillator with good frequency stability, a phase shifter followed by shaping circuits may be used to obtain a second pulse with accurately controlled delay relative to an unshifted trigger pulse. This system has been used in radar equipment to measure range accurately.

Repetition rate may be determined by a free-running relaxation oscillator such as a multivibrator, blocking oscillator, or thyratron. The repetition rate is then usually more subject to drift than if an L-C oscillator is used, but a suitable trigger pulse and a large range of repetition rates are economically obtained. Figure 13 shows a relaxation oscillator

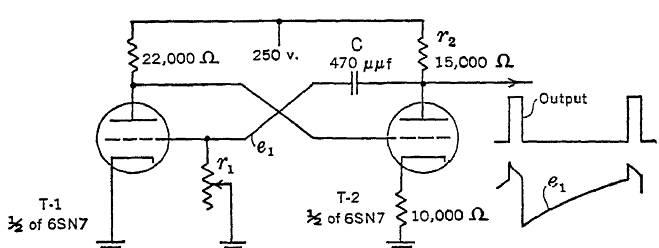


FIG. 13. Repetition Rate Determined by an Exponential Wave Form

providing trigger pulses of fixed duration as the repetition rate is changed. The pulse duration is determined by the exponential charging of capacitor C through r_2 and the low grid-to-cathode resistance of tube T-1. After each pulse the grid of tube T-1 is driven negative for a period determined by the exponential discharging of capacitor C through r_1 and the relatively small resistance of r_2 and tube T-2. The period between pulses is approximately proportional to the resistance r_1.

In some applications a delay line may be used to control repetition rate and simultaneously supply synchronizing pulses at desired times in each period. Thus, in Fig. 14, the blocking oscillator generates pulses which after traversing the delay line are applied to the trigger tube and initiate new pulses.

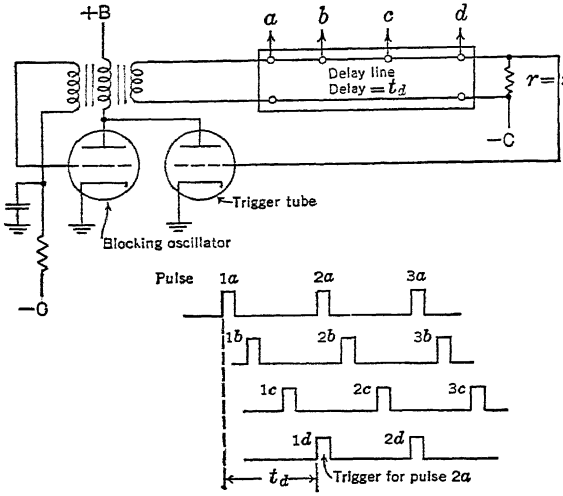


FIG. 14. Repetition Rate Determined by a Rectangular Wave Form

Duration. Pulses having a required duration may be generated by relaxation circuits or by pulse-forming lines and thyratrons, as discussed in article 15, "Pulse Modulation of an Oscillator." Alternatively, sinusoidal or other wave forms may be passed through pulse shaping circuits (see above) to produce pulses of the required duration.

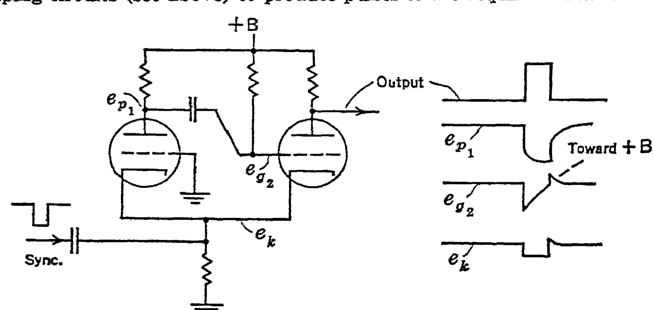


FIG. 15. Pulse Duration Determined by an Exponential Wave Form

A cathode-coupled one-pulse relaxation circuit is shown in Fig. 15 to illustrate control of pulse duration by exponential voltage change.

Blocking oscillators deliver maximum power during the pulse and therefore are adapted for high-power pulse equipment. Nominal control of pulse duration is obtained by means of grid capacitors and control of transformer inductance, in the circuit of Fig. 12. Precise duration and better pulse shape are obtained by means of a delay line as in Fig. 16. The transformer inductance is made large; the feedback, the line impedance, and the line terminating resistance are so chosen that the oscillator is shut off when the wave of grid

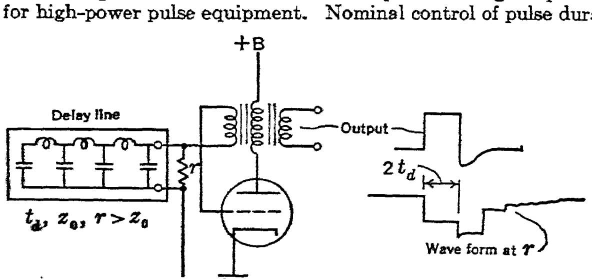
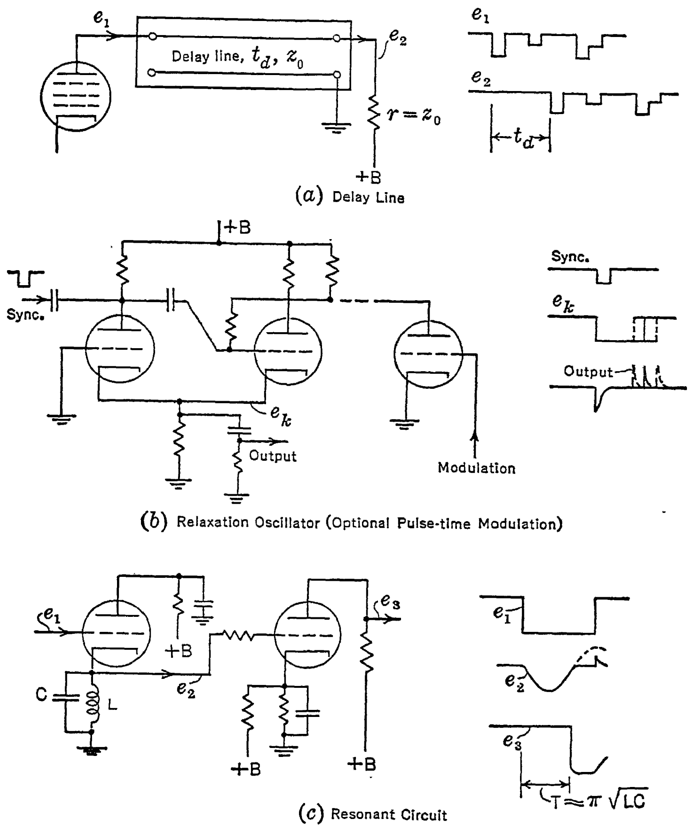


FIG. 16. Pulse Duration Determined by a Rectangular Wave Form

current reflected from the open end of the line reaches the terminating resistor. The pulse duration is then determined mainly by the round trip delay of the line and is little affected by variations in other circuit elements.

Delay. A complex wave may be delayed by passage through a delay line while maintaining substantially the same wave form. If it is only required that a delayed pulse be produced, the initiating pulse may trigger a relaxation oscillator, and the trailing edge of the relaxation pulse may be used after differentiation as the delayed pulse. The delay of the trailing edge of the relaxation pulse may be varied in accordance with a desired



(*a*) Delay Line

(*b*) Relaxation Oscillator (Optional Pulse-time Modulation)

(*c*) Resonant Circuit

Fig. 17. Circuits for Delaying Pulses

modulation for obtaining pulse-time modulation in a communication system. Special cathode-ray tubes have been developed to facilitate time modulation of the many time-sharing pulse channels in pulse-time multiplex systems. In another method, a resonant circuit is keyed on or suddenly shocked, and a delayed pulse is produced after a fraction of a cycle of oscillation. Typical circuits using (*a*) the delay line, (*b*) the relaxation oscillator, and (*c*) the resonant circuit are shown in Fig. 17.

The delay line can accept pulses having random spacing such as might be encountered in the output of a receiver. The other circuits illustrated must pass through a complete delay and recovery cycle between pulses and therefore are best used with isolated pulses.

15. PULSE MODULATION OF AN OSCILLATOR

An oscillator may be made to operate in pulses by including in its grid circuit a capacitor which charges during the oscillation and blocks or stops the oscillation. This is one form of the blocking oscillator. Oscillations start again after the capacitor has discharged sufficiently. The pulse recurrence rate and the pulse duration are approximately controlled by the time constants in the circuit, but only with limited stability.

The recurrence rate of a blocking oscillator may be stabilized by the use of synchronizing circuits to start tube conduction sooner after each pulse than would otherwise happen. The pulse duration may be stabilized by making a time-delay network shut off the oscillator.

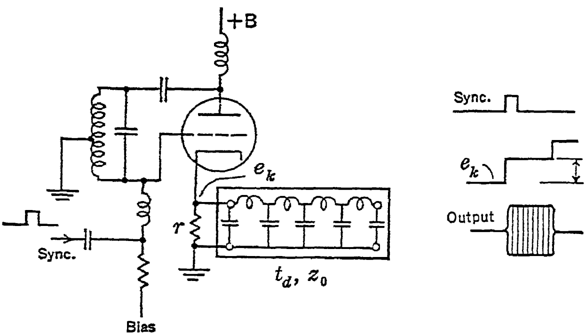


FIG. 18. Pulse-modulated Carrier-frequency Oscillator

The oscillator in Fig. 18 has provision for synchronization by a positive pulse applied to the grid. A delay line in the cathode circuit shuts off the oscillator after a time interval of twice the line delay. The discharging resistor r is usually many times as large as Z_0, the wave impedance of the line. At the start of the pulse the cathode current flows through an impedance approximately equal to Z_0, and develops a voltage wave E which travels through the line. The wave is reflected with the same polarity by the open far end and returns to increase the voltage at the cathode to approximately twice the initial voltage, stopping the oscillator.

Pulse modulation may also be accomplished by grid or plate modulation of an oscillator or amplifier. If high power is required, plate modulation of an oscillator is generally most economical. With plate modulation, the high voltage is impressed on the oscillator tube only during the pulses. The design of the oscillator tube is thereby made easier, or conversely the permissible maximum plate voltage on a given tube design may be increased far beyond the direct-voltage rating.

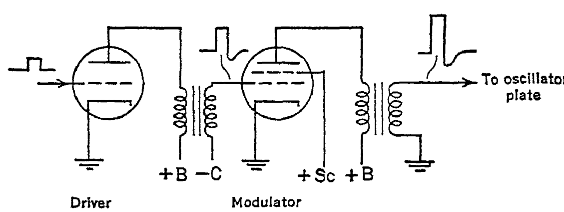


FIG. 19. Vacuum-tube Modulator

A vacuum-tube modulator is shown in Fig. 19. The modulator tube, cut off between pulses, is driven hard so that it passes high current through low impedance during the pulses. The modulator acts as a switch connecting the load to the B supply only during the pulses.

A thyratron pulse modulator is shown in Fig. 20. The artificial line is charged through a high resistance r_c between pulses. By the use of the line instead of a single capacitor, a

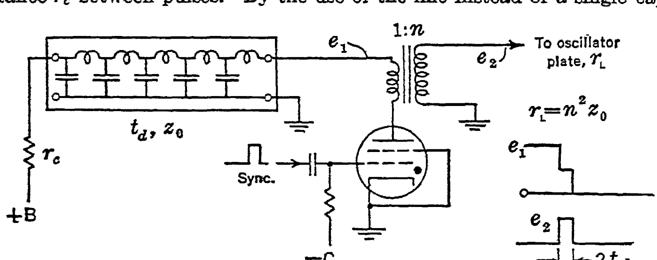


FIG. 20. Thyratron Modulator

rectangular pulse may be obtained rather than an exponentially decaying pulse. The thyratron discharges the line into the load and must deionize before the line can recharge. Deionization is expedited by making the thyratron resistance plus the transformed load resistance somewhat less than the line impedance so that the thyratron plate is driven slightly negative just after the pulse.

For high-voltage applications, the preferred form of artificial line is the "Guillemin line," shown in Fig. 29(*e*) below. Also, the thyratron may be replaced by a spark gap. Higher efficiency and faster charging of the line may be obtained by replacing the charging resistor r_c by an inductor. The line is then charged in a half-cycle of a charging oscillation, between pulses, to approximately twice the power supply voltage. A diode may be connected in series with the inductor to prevent discharge back into the power supply.

Pulse modulation imposes on an oscillator severe requirements of fast starting and stopping. The oscillator must ordinarily start from the background noise in the circuit and build up to maximum amplitude in a time interval which is much shorter than the pulse duration. Random variations in the noise amplitude cause random variations in starting time, termed "jitter."

To explain the starting and stopping time, the essential elements of the oscillator are indicated in Fig. 21. They are the resonator *CL* and the conductance *g*. During the starting time of a pulse, the net conductance is made negative by feedback; during the stopping time it is positive as determined by the damping of the circuit augmented by its useful load.

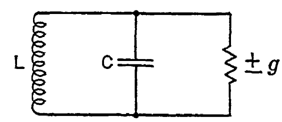


FIG. 21. Equivalent Circuit of an Oscillator

The starting or stopping time constant of oscillations is $2C/g$ The starting time, for the oscillations to build up by regeneration from the noise level, is of the order of 20 times the time constant determined by the negative conductance, and the jitter is somewhat less than this time constant. The stopping time is about equal to the stopping time constant, since the oscillation is merely damped from its peak value after the regeneration is cut off. These effects are shown in Fig. 8, p. 9-06.

16. MODULATING THE CHARACTERISTICS OF PULSES

Preceding sections of this chapter have indicated ways of forming pulses and controlling their characteristics. Coding of pulses for identification of the source or for transmission of information may be accomplished by mechanical switching or by electronic switching or modulation. Pulse duration may be changed by connecting or disconnecting sections in a delay line used to control a blocking oscillator. Similarly, the constants of a shaping circuit or relaxation circuit may be switched.

The circuit shown in Fig. 17(*b*) permits electronic modulation of pulse width such as might be used in a communication system. Differentiation of the output pulse, as shown in the figure, yields a pulse having electronically modulated phase relative to a base or marker pulse. Pulse phase or time may also be modulated at an audio rate by addition of audio voltage to linear sawtooth timing pulses as shown in Fig. 22. No plate current

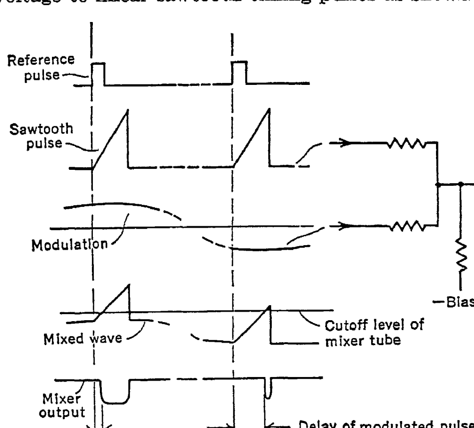


FIG. 22. Pulse-time Modulator

flows until the sum of the audio, the sawtooth, and the bias voltages exceeds the cutoff voltage of the mixer tube. The timing of the leading edge of the resulting plate-current pulse varies in direct proportion to the instantaneous value of the audio voltage. The mixer output pulse may be differentiated and shaped to obtain the desired wave form for pulse modulation of a carrier wave.

A group of pulses may be obtained by shaping a sine wave which has the proper periodicity, to obtain trigger pulses for a relaxation circuit. The sine wave is initiated and ended by a gate pulse long enough to permit generation of the required number of pulses. Figure 23 shows one such arrangement. Triode A is conducting normally. A negative

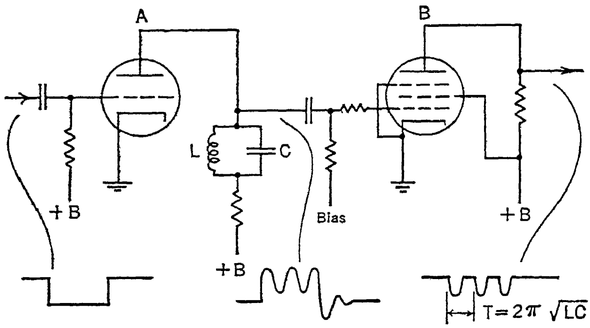


FIG. 23. Multiple-pulse Generator

gate pulse cuts off the triode, developing a positive gate and a free oscillation to drive the grid of the pentode B. At the end of the gate pulse the oscillations are quickly damped by the plate resistance of the triode. The number of pulses may be changed by switching the gate pulse duration.

Figure 24 shows the elements of a two-pulse generator using a delay line. A modification of this arrangement in which the pulse spacing would correspond to altitude has been proposed for use in air navigation (Lanac). An initiating pulse passes through a delay

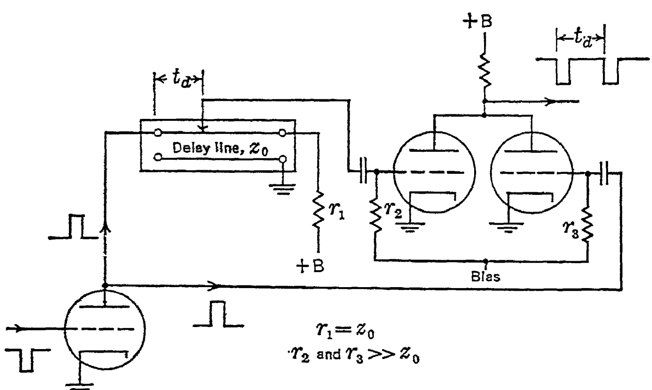


FIG. 24. Double-pulse Generator

line and is absorbed by the resistor terminating the line. The line is tapped to permit application of the pulse, after the required delay, to the grid of a tube. The direct and delayed pulses are mixed to produce a pair of pulses with variable spacing.

17. PULSE DETECTORS

Pulses are distinguished from sinusoidal and steady waves by rapid changes in amplitude, large ratio of peak to average value, small duty cycle, and wide frequency spectrum. The detection of pulses requires detecting means responsive to the special characteristics of pulses, usually to their time boundaries as distinguished from their peak amplitudes.

Bandwidth requirements have been discussed earlier in this section. Pulse amplifiers may include filters or coupling systems which favor pulses having certain characteristics, require more pulses in a group. The timing of two or more pulses in a group may be used for altitude coding in an air navigation system (Lanac) as mentioned previously.

Phase or time modulation of pulses in a communication system may be converted to audio signals by a corresponding type of detection. Phase modulation may be detected by providing reference pulses of uniform periodicity and using these to form gate pulses of suitable phase which the phase-modulated pulses overlap more or less in accordance with the modulation. If the gate and the phase-modulated pulses are applied to a coincidence mixer, plate current will flow during the interval of pulse overlap. The plate current, averaged over a few pulses, is then modulated in accordance with the phase modulation of the pulses. Special beam-deflection tubes have been developed for phase modulation and synchronous detection of many time-sharing channels in a pulsed multiplex system.

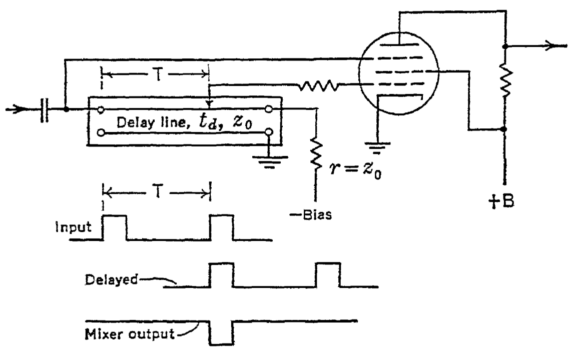


FIG. 27. Double-pulse Decoder

18. VACUUM TUBES

Most vacuum tubes, at the date of writing, have not been rated for pulse applications (see Section 4, especially pulse tubes such as magnetrons and thyratrons). However, many vacuum tubes have been used in pulse applications after rough estimation of their pulse capabilities from their ordinary ratings. The principal factors are peak and average cathode current, peak and average electrode voltages, electrode power dissipations, and tube life.

If a tube has class C amplifier ratings, the cathode current for long pulses may safely be 4 or 5 times the class C average-current rating. The average current should not exceed the class C average-current rating. For pulses a few microseconds in duration, peak currents of about 50 times the class C average-current rating may be taken from oxide-coated cathodes, with tube life commonly exceeding 500 hours. Pulse current densities of 10 to 20 amp per sq cm are used with specially processed oxide-coated cathodes. The cathodes must be heated to their operating temperature before large currents are permitted.

The ordinary peak-voltage ratings of tubes may often be safely exceeded in pulse applications. Insulation breakdown inside the tube, electrolysis of the glass, or poor plate-current cutoff characteristics may fix the permissible peak voltage. The peak plate voltage in class C applications is about twice the average plate voltage; this is usually a justification for permitting pulse peak voltages twice the rated average plate voltage. If the plate voltage is applied only during pulses, the sparkover voltage may set the limit at a much higher voltage.

Beam power tubes in pulse modulators are operated at high screen voltage in order to reduce the driving power required. The plate voltage is usually low during pulses of current flow, so that average plate power dissipation is low. Screen dissipation is high, however, and often limits average power. Permissible limits of plate and screen dissipation may be reduced as peak-voltage ratings are increased.

The space-charge effects which suppress secondary emission from the plate in beam tubes also tend to limit the flow of pulse current at low plate voltage. The usually desired properties of beam tubes are even more important for pulse amplifiers of very high power used as modulators. Secondary emission from the control grid is minimized to avoid loss of control caused by reverse grid current and to prevent parasitic oscillation.

A number of thyratrons specifically intended for pulse modulation are available. Miniature types deliver pulses of peak power of kilowatts; larger types deliver megawatts. Fast deionization is obtained in tubes filled with the lighter gases such as hydrogen and helium, slower deionization with argon and xenon. Cathode current densities are of the same order as in high-vacuum tubes but with much less voltage drop between plate and cathode. The internal voltage drop is high at the start of ionization, very rapidly decreasing as ionization reaches saturation. A maximum rate of rise of current is specified to limit the plate dissipation and the ion bombardment of the cathode at the start of the pulse. The rate of rise of current is usually limited by some inductance in series with the plate of the thyratron.

It is worthy of mention that, at the high currents used in pulse work, the transconductance of vacuum tubes is greater than usual, and transit time is reduced by the corresponding high voltages. The performance of a tube in a pulsed oscillator is therefore often better than its continuous-wave performance at high carrier frequencies subject to transit-time effects. In fact, some types will oscillate efficiently in high-power pulses at frequencies so high that continuous oscillation is impossible at their maximum d-c ratings.

19. PULSE TRANSFORMERS

The problem of designing a pulse transformer is generally similar to that of designing an audio-frequency transformer but for higher-frequency components with corresponding

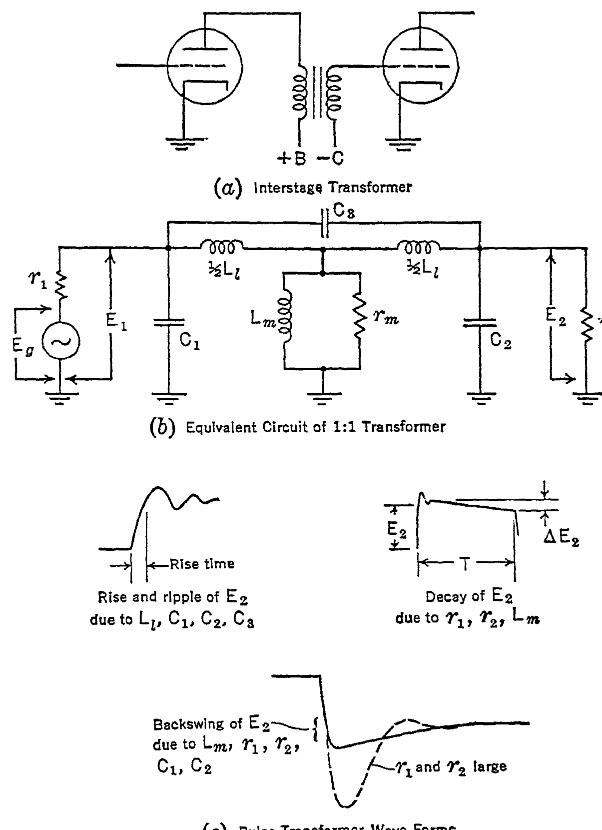


Fig. 28. Pulse Transformers

reduction in size. Maximum mutual inductance, minimum leakage inductance, and minimum incidental capacitance are desired. The ratio of maximum to minimum essential frequency components involved in faithfully transforming a single pulse wave form may be 100 to 1. If the pulse duration may vary by 10 to 1, the frequency ratio may be 1000 to 1. The requirements for high-power modulator transformers are usually much less severe in frequency ratio.

A rough estimate of transformer specifications may be obtained by analysis of the circuit of Fig. 28(a) and the equivalent circuit shown in Fig. 28(b). A turns ratio of 1 to 1 is assumed. The mutual inductance L_m may be neglected in estimating the rise time of the pulse. If r_1 and C_1 are small, the rise time is roughly $\sqrt{L_1C_2}$ or L_1/r_2, whichever is greater. The magnetizing current at the end of the pulse is

$$I_m = \frac{E_1T}{L_m} + \frac{E_1}{R_m} \tag{1}$$

if the pulse duration is T. The decrease in voltage during the pulse and the backswing at the trailing edge, both of which are caused by the magnetizing current, are roughly in the ratio

$$\frac{\Delta E_2}{E_2} = \frac{T}{L_m}\frac{r_1r_2}{r_1 + r_2} \tag{2}$$

if r_1 and r_2 are small and linear. If r_1 and r_2 become large on reversal of polarity, the backswing is determined by I_m, L_m, $C_1 + C_2$, C_3, and r_m. r_m represents the apparent shunt resistance of eddy currents and hysteresis in the core loss.

To reduce eddy currents or skin effect in the core, very thin laminations are used for pulse transformers. Permalloy and silicon-steel tapes from 0.0001 to 0.003 in. in thickness are wound into toroidal cores which are sliced for assembly with coils. Reduction of skin effect in the core permits greater ratios of mutual to leakage inductance by increasing the depth of penetration and the resulting effective permeability of the core volume. The available flux swing in the core may sometimes be increased by the use of reverse-magnetizing direct current. Because pulse wave forms are usually unidirectional, and the core has some retentivity, saturation may otherwise occur with a small flux swing per pulse. Some core materials especially suited for pulse applications are listed below.

Because of restrictions on leakage inductance, high-voltage pulse transformers are difficult to insulate. Impregnation with oil or polymerized resins to eliminate air spaces and moisture are effective in preventing corona and raising the breakdown voltage.

Table 1. Core Materials for Pulse Applications

Name and Manufacturer	*Description*
Hypersil Westinghouse Electric Corp.	Wound silicon-steel tape: grain-oriented 0.002 in' thick tape for high quality; plain 0.003 in. tape for less critical application.
Permalloy Bell Telephone Laboratories	Wound permalloy tape: from 0.0001 to 0.002 in. thick; several different alloys.
Silicon Nicalloy General Electric Co.	Conventional laminations: type B9W4A.
Sinimax and Monomax Allegheny Ludlum Steel Corp.	Conventional laminations: several different alloys (permalloys).

20. DELAY LINES *

A delay line is a network used for storing the energy of a pulse pattern and delivering out the same energy at a later time. (See Section 5; Networks, Lines, Transients.) The simplest form is a transmission line, in which case the delay is determined by the distance along the line at a velocity somewhat less than the velocity of light. While the transmission line or wave guide may give the best performance, it requires a length too great for most purposes, so more concentrated forms have been devised to save space.

Delay lines are used for two general purposes: one is the relative timing of different operations; the other is the delayed reproduction of a pulse pattern. The former may be tolerant of distortion in the delay process and can even be accomplished by other kinds of timing circuits such as relaxation oscillators. The delayed reproduction of a pulse pattern, however, places the most severe requirements on a delay network, especially against amplitude and phase distortion. For either purpose, precision of timing may be a requisite which involves directly the stability of the network and indirectly the fidelity of reproduction and the stability of associated circuits.

Figure 29 shows some typical concentrated delay networks. They rely on coils for concentrating the inductance and on dielectric for concentrating the capacitance. Further concentration of inductance may be obtained by iron-dust cores for the coils.

The continuous coil with capacitive loading is shown in Fig. 29(a). In a typical form, a coil of fine enameled wire is wound on a flexible core of insulation material having such small diameter that it can be coiled. The winding is wrapped with thin dielectric tape for capacitive loading, and the other conductor is provided by a braided sheath of fine

* This article by Harold A. Wheeler.

enameled wire. The sheath acts as a capacitive shield but not as an inductive shield, because the latter would destroy the desired inductance of the coiled inner conductor. The entire line is protected by a covering of insulation.

In a continuous coil, the mutual inductance along its length has the effect of decreasing the effective inductance at higher frequencies relative to that of lower frequencies, which causes phase distortion. In Fig. 30, the curved dotted line (*a*) shows this effect. Uniform delay, free of distortion, requires a straight phase curve illustrated by the solid line (*b*). It has been found possible to approximate this ideal by introducing substantial distributed capacitance in parallel with the series inductance, as is also shown in Fig. 29(*a*). This causes the effective inductance to increase with frequency and can be designed to compensate for the opposite effect of mutual inductance along the coil. Without such compensation, the effect may be reduced by using a coil diameter very small as compared with its length, which tends, however, to defeat the aim of concentration.

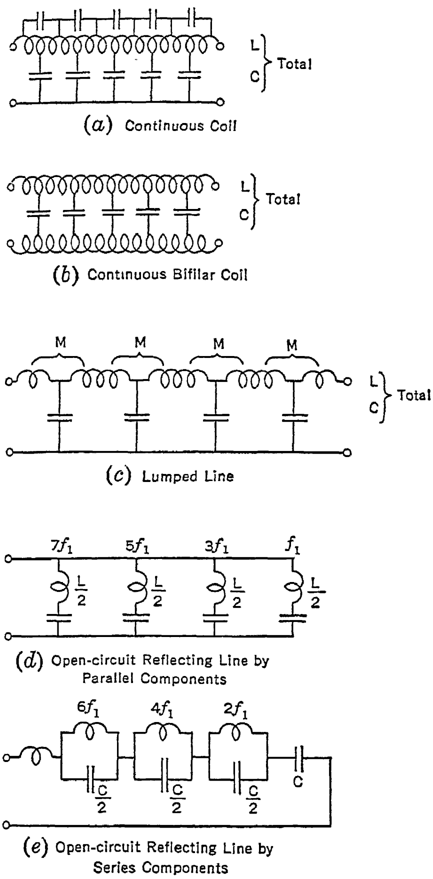


FIG. 29. Delay Lines

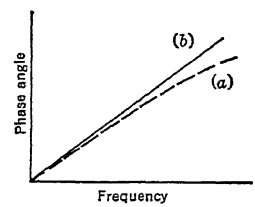


FIG. 30. Phase Distortion of Continuous Coil

The low-frequency delay (initial phase slope) of a delay line is

$$T = \sqrt{CL} \tag{3}$$

in terms of the total shunt capacitance C and total series inductance L (including mutual inductance). The corresponding wave impedance (characteristic impedance or image impedance) is

$$Z_0 = \sqrt{L/C} \tag{4}$$

Any of the following sets of units may be specified, the last (long used by Dr. Alan Hazeltine) being generally most convenient:

T	seconds	μs	μs
C	farads	μf	mμf
L	henrys	μh	mh
Z_0	ohms	ohms	kilohms

As an example, a rather long coiled line of 1 mμf and 1 mh has a delay of 1 μs and a wave impedance of 1 kilohm.

Another form of coiled line, shown in Fig. 29(*b*), is made of two coils wound in opposite screw-directions on the same core. One coil is wound in one direction on the core, then the other is wound on top in the opposite direction. The insulation may be only the enamel on the wires, or it may also include some thin layers of dielectric. This method gives the inductance of a two-layer coil with convenient capacitive loading. It is useful as a balanced four-terminal circuit or as a two-terminal reflecting line with far end on open circuit or short circuit.

The low-pass filter of Fig. 29(*c*) is a lumped delay line. Its useful bandwidth is somewhat less than its cutoff frequency and decreases further with more sections. The constant-k type (without mutual inductance) has concave phase distortion as shown in Fig. 31(a). By converting to an m-derived type, by the addition of mutual inductance M between adjacent inductors, it is possible to approach the ideal linear phase (b) to the approximation indicated by curve (c). For a line of many sections, the optimum value of the design parameter m is $\sqrt{3/2} = 1.22$. The number of sections required is

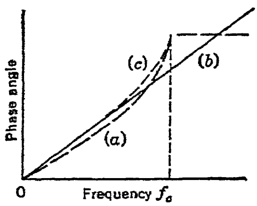


FIG. 31. Phase Distortion of Lumped Line

$$N = \frac{\pi}{m} f_c T \tag{5}$$

in terms of the cutoff frequency f_c (Mc) and the delay T (μs). As an example, a pulse 1 μs wide can be delayed 2 μs by a network of $f_c = 1$ Mc and $N = 5$ sections.

In a concentrated delay line with phase correction, especially if the delay is many times the permissible widening of the pulse, the attenuation increasing with frequency is likely to impose the practical limitation on the useful bandwidth. Therefore all losses must be minimized to secure the optimum design in limited space.

A reflecting line is often useful and has the advantage of a round trip which doubles the delay. The voltage polarity of reflection is direct or reverse, depending on whether the far end of the line is an open or short circuit. Ordinarily the near end is terminated with a resistance load matching the line impedance to preclude multiple reflection.

A reflecting line becomes a two-terminal network whose essential properties are expressed by the variation of impedance with frequency. An ideal delay line of uniform delay and no losses has a pure reactance as shown in Fig. 32 if the far end is on open circuit. The alternate zeros and poles are at odd and even multiples of the fundamental frequency:

$$f_1 = \frac{1}{4\sqrt{CL}} \tag{6}$$

At this frequency, the length is 1/4 wavelength in the line.

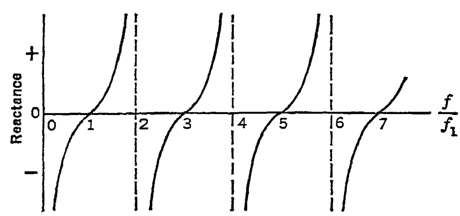


FIG. 32. Impedance of Open-circuit Reflecting Line

The impedance of a reflecting line can be duplicated over a limited frequency bandwidth by other two-terminal impedance networks which sometimes have advantages. Figure 29(d) presents the reactance pattern of Fig. 32 by a parallel connection of several series-tuned circuits having equal values of inductance but such values of capacitance as to resonate at odd multiples of the fundamental frequency. Figure 29(e) presents the same reactance pattern by a series connection of parallel-resonant circuits having equal values of capacitance and resonated at even multiples. This last network has a great advantage in pulse discharging functions (as proposed by E. A. Guillemin) because only the isolated capacitor C has to take the high-voltage charge preparatory to the discharge; the other capacitors can have much lower voltage ratings.

When a charged delay line is discharged into a matched load, the output is a pulse whose width is the round-trip delay $2T$, as shown in Fig. 33. The pulse starts suddenly by the discharge of energy near the terminals, but the trailing edge is sloping because the delayed energy is subject to the distortion in the line.

FIG. 33. Pulse Discharge of a Line into a Matched Load

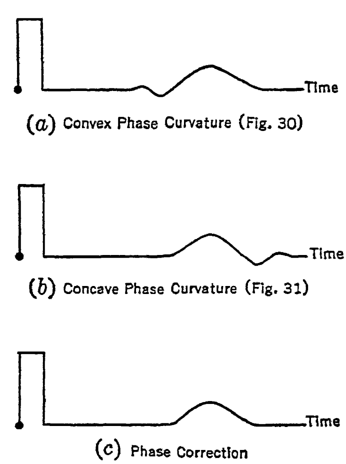


FIG. 34. Delay of a Pulse by a Reflecting Line on Open Circuit, Showing the Effects of Distortion

Figure 34 shows the distortion of a pulse subjected to round-trip delay in a reflecting line on open circuit. In every case the pulse is decreased in area by the attenuation and is widened by the amplitude distortion limiting the frequency bandwidth. The effects of phase distortion can be estimated by the method of paired echoes. The leading transient (*a*) and the trailing transient (*b*) are characteristic of convex and concave phase curvature, shown in Figs. 30 and 31. The former occurs in continuous coiled lines and the latter in lumped lines. In either case, phase correction (Fig. 34*c*) removes the transient oscillation, since the oscillation is caused by subnormal or abnormal delay of the higher-frequency components of energy in the pulse.

FIG. 35. Delay and Reversal of a Pulse by a Reflecting Line on Short Circuit

If the direct and reflected pulses have to be separated, a delay line on short circuit (or an equivalent two-terminal network) is used to secure the result shown in Fig. 35. Since the respective pulses are of opposite polarity, one or the other can be selected by a rectifier.

BIBLIOGRAPHY

Note. Under each heading, the references are listed chronologically (except for the "PB" items, whose dates of publication do not appear in their titles). The "PB" items are issued by Office of Technical Services, U. S. Department of Commerce, Washington 25, D. C.

Asterisks (*) denote some of the principal references or further bibliographies on each subject.

Radar

*1. Navy Dept., Radar System Fundamentals, *Navships* 900,017 April 1944.
*2. Navy Dept., Radar Electronic Fundamentals, *Navships* 900,016, June 1944.
3. D. G. Fink, The Radar Equation, *Electronics*, Vol. 18, No. 4, 92–94 (April 1945).
4. W. E. Moulic, Operational Elements of a Radar System, *Electronic Industries*, Vol. 4, No. 5, 76–80, 225, 226 (May 1945).
5. Jordan McQuay, Practical Radar, *Radio News:* Part I, Vol. 33, No. 6, 29 (June 1945); Part 2, Vol. 34, No. 1, 38 (July 1945); Part 3, Vol. 34, No. 2, 39 (August 1945); Part 4, Vol. 34, No. 3, 40 (September 1945); Part 5, Vol. 34, No. 4, 44 (October 1945).
*6. R. B. Colton, Radar in the United States Army, *Proc. I.R.E.*, Vol. 33, 740–753 (November 1945).
7. Report on Wartime Electronic Developments, *Electronics*, Vol. 18, No. 11, 92–93 (November 1945).
*8. Radar Specifications, *Electronics*, Vol. 18, No. 11, 116–119 (November 1945).
*9. The SCR-584 Radar, *Electronics:* Part I, Vol. 18, No. 11, 104–109 (November 1945); Part II, Vol. 18, No. 12, 104–109 (December 1945).
10. C. W. Watson, Ground-controlled Approach for Aircraft, *Electronics*, Vol. 18, No. 11, 112–115 (November 1945).
11. H. A. Straus, L. J. Rueger, C. A. Wert, S. J. Reisman, M. Taylor, R. J. Davis, and J. H. Taylor, The MPG-1 Radar, *Electronics:* Part I, Vol. 18, No. 12, 92–97 (December 1945); Part II, Vol. 19, No. 11, 110–117 (January 1946).
12. W. C. Hendricks, Lightweight Radar for Early Warning, *Communications*, Vol. 26, 54 (January 1946).
*13. W. C. Tinus and W. H. C. Higgins, Early Fire-control Radars for Naval Vessels, *Bell Sys. Tech. J.*, Vol. 25, 1–47 (January 1946).
*14. H. A. Zahl and J. W. Marchetti, Radar on 50 Centimeters, *Electronics*, Vol. 19: Part 1, No. 1, 98–104 (January 1946); Part 2, No. 2, 98–103 (February 1946).
15. W. T. Spicer, The GCA Landing System, *Bendix Radio Engineer*, January 1946. (Radar on 3000 Mc or 10,000 Mc with PPI display.)
16. H. V. Hermansen, *Bendix Radio Engineer*, January 1946.
17. A Tool for Traffic Control—Telerad, *Air Transport*, Vol. 4 /No. 1, 73–78 (January 1946)
18. I. F. Byrnes, Merchant Marine Radar, *R.C.A. Rev.*, Vol. 7, 54–66 (March 1946).
19. Jack Mofenson, Radar Echoes from the Moon, *Electronics*, Vol. 19, No. 4, 92–98 (April 1946).
*20. D. A. Quarles, Radar Systems Considerations, *Elec. Eng.*, Vol. 65, *Trans.*, pp. 209–215 (April 1946).
*21. G. V. Holdam, S. McGrath, and A. D. Cole, Radar for Blind Bombing, *Electronics*, Vol. 19: Part I, No. 5, 138–143 (May 1946); Part II, No. 6, 142–149 (June 1946).
*22. O. P. Ferrell, London Radiolocation Convention, *Radio*, Vol. 30, No. 5, 26, 28, 47, 48 (May 1946).
23. L. N. Ridenour, Radar in War and Peace, *Elec. Eng.*, Vol. 65, 202–207 (May 1946).
24. R. C. Jensen and R. A. Arnett, Air-borne Radar for Navigation and Obstacle Detection, *Elec. Eng.*, Vol. 65, No. 5, 307–313 (May 1946).
25. H. Busignies, P. R. Adams, and R. I. Colin, Aerial Navigation and Traffic Control with Navaglobe, Navar, Navaglide and Navascreen, *Elec. Communication*, Vol. 23, No. 2, 113–143 (June 1946).
26. E. I. Green, H. J. Fisher, and J. G. Ferguson, Techniques and Facilities for Microwave Radar Testing, *Bell Sys. Tech. J.*, Vol. 25, No. 3, 435–482 (July 1946).
27. Thomas Grover and E. C. Kluender, The Electronic Navigator, *Communications*, Vol. 26, No. 8, 30, 36–39 (August 1946).
*28. E. G. Schneider, Radar, *Proc. I.R.E.*, Vol. 34, 528–578 (August 1946).
29. L. R. Quarles and W. M. Breazeale, Factors Affecting the Range of Radar Sets, *Elec. Eng.*, Vol. 65, *Trans.*, pp. 546–548 (August–September 1946).
*30. L. V. Berkner, Naval Airborne Radar, *Proc. I.R.E.*, Vol. 34, 671–706 (September 1946).
31. E. D. Hart, Navigational Radar in Merchant Ships, *Electronic Eng.*, Vol. 18, 265–267 (September 1946).
32. J. H. Cook, Airborne Search Radar, *Bell Lab. Rec.*, Vol. 24, 321–325 (September 1946).

33. M. L. Lawrence, A 100-kw Portable Radar Transmitter, *Communications*, September 1946, pp. 30, 32, 33.
34. C. B. Barnes, Radar Carrier-based Planes, *Electronics*, Vol. 19, No. 10, 100–105 (October 1946).
35. H. B. Brooks, Weather Forecasting, *Electronics*, Vol. 19, No. 10, 84–87 (October 1946).
36. M.I.T. Radar School Staff, *Principles of Radar*, 2nd Ed. McGraw-Hill (1946).
37. L. H. Lynn and O. H. Winn, "Marine Radar for Peacetime Use," *Elec. Eng.*, *Transactions*, Vol. 65, 271 (May 1946).
38. PB18042, Navy Dept., Radar SU-1 with PPI display.
39. PB19248, War Dept., Radar APS-4A on 9400 Mc.
40. PB20810, Navy Dept., Radar APS-4.
41. PB20812, War Dept., Radar APS-15 training course.
42. PB20813, Navy Dept., Radar APS-3, on 9400 Mc.
43. PB22400, Navy Dept., Radar CXAM-1 on 195 Mc.
44. PB23435, War Dept., Radar TPS-3.
45. PB23438, War Dept., Radar SCR-682-A.
46. PB23447, War Dept., Radar SCR-682-A.
47. PB24913, Navy Dept., Radar SC-3.
48. PB24933, Navy Dept., Radar SR with A-type and PPI displays.
49. PB24936, Navy Dept., Radar SJ-1 with PPI display.
50. PB24938, Navy Dept., Radar SA-3 on 200 Mc.

Radio Pulse Altimeters

61. Heinrich Lowry, Electric Proof and Measuring of the Distance of Electrically Conductive Bodies, U. S. Pat. 1,585,591, July 17, 1923–May 18, 1926.
62. Gregory Breit and M. A. Tuve, A Test of the Existence of the Conducting Layer, *Phys. Rev.*, Vol. 28, 554–575 (September 1926).
63. R. W. Hart, Measuring Distance, U. S. Pat. 1,924,156, May 19, 1930–Aug. 29, 1933.
64. Ezekiel Wolf, Measuring Distance, U. S. Pat. 1,924,174, May 19, 1930–Aug. 29, 1933.
65. D. L. Plaistowe, Radio Apparatus for Detecting Aircraft, U. S. Pat. 2,207,267, July 11, 1938–July 9, 1940.
66. Joseph Lyman, Radio Absolute Altimeter, U. S. Pat. 2,227,598, July 3, 1937–Jan. 7, 1941.
*67. Albert Goldman, Pulse-type Radio Altimeter, *Electronics*, Vol. 19, No. 6, 116–119 (June 1946).
68. P. G. Sulzer, Ionosphere Measuring Equipment, *Electronics*, Vol. 19, No. 7, 137–141 (July 1946).

Navigation and Identification by Pulses

*71. Editor (D. G. Fink), The Loran System, Part I, *Electronics*, Vol. 18, No. 11, 94–99 (November 1945). Loran Receiver-indicator, Vol. 18, No. 12, 110–115 (December 1945). Loran Transmitting Stations, Vol. 19, No. 3, 109–115 (March 1946).
72. Editor, Fundamentals of Radar—Pulse Methods Applied to Navigation, *Wireless World*, Vol. 52, 23–26 (January 1946).
73. Editor, Fundamentals of Radar. 5. Beacons Employing Pulse Technique, *Wireless World*, Vol. 52, 55–56 (February 1946).
74. D. Davidson, Loran Indicator Circuit Operation, *Elec. Ind.*, Vol. 5, No. 3, 84–93, 126, 128, 130, 132 (March 1946).
*75. S. W. Seeley, Shoran Precision Radar, *Elec. Eng.*, Vol. 65, *Trans.*, pp. 232–240 (April 1946).
*76. J. A. Pierce, An Introduction to Loran, *Proc. I.R.E.*, Vol. 34, 216–234 (May 1946).
*77. Hazeltine Electronics Corp., *Lanac Air and Marine Navigation* (1946).
78. PB20811, War. Dept., Radar Transpondor Beacon APN-7 on 3300 Mc.
79. PB23325, P. J. Herbst, Ultra Portable Racon.
*80. PB23434, War. Dept., Radar RC-145-A.
81. PB23448, War Dept., Radar RC-150B, RC-151, etc.
82. PB24920, Navy Dept., Radar CPN-3.
83. PB24922, Navy Dept., Radar BQ.

Sonar

*91. R. J. Evans, Echo Ranging Sonar, *Electronics*, Vol. 19, No. 8, 88–93 (August 1946).
*92. G. B. Shaw, Echo Depth Sounder for Shallow Water, *Electronics*, Vol. 19, No. 9, 88–92 (September 1946).
93. Thomas Roddam, Radar in Nature, *Wireless World*, Vol. 52, 286–288 (September 1946).
94. PB19955, Navy Dept., Sonar QBE, ABE, etc.
95. PB25497, O. R. Smith, Fish Schools—Location by Sonic Equipment.

Pulse Techniques and Measurements

*101. D. G. Fink, *Microwave Radar*, Vol. 1, *Theory and Practice of Pulsed Circuits* (July 1942).
*102. J. G. Brainerd, *Ultra-high-frequency Techniques*. Van Nostrand (1942). Chapter 4, trigger circuits, pulse-sharpening circuits and oscillators, pp. 168–207.
*103. Navy Dept., Microwave Techniques, *Navships* 900028 (June 1944).
104. Allan Easton, Pulse Response of Diode Voltmeters, *Electronics*, Vol. 19, No. 1, 146–149 (January 1946).
105. Allan Easton, Measuring Pulse Characteristics, *Electronics*, Vol. 19, No. 2, 150–154 (February 1946).
*106. G. M. Lee, A Three-beam Oscillograph for Recording at Frequencies up to 10,000 Mc, *Proc. I.R.E.*, Vol. 34, 121W–127W (March 1946).
*107. E. I. Green, H. J. Fisher, and J. G. Ferguson, Techniques and Facilities for Microwave Radar Testing, *Elec. Eng.*, Vol. 65, 274–290 (May 1946).
*108. M.I.T. Radar School Staff, *Principles of Radar*, 2nd Ed. McGraw-Hill (1946).
109. PB21937, War Dept., Radar Signal Generator TS-155B/UP on 2700–2900 Mc.
110. PB23432, War. Dept., Range Calibrator I-146.
111. PB23625, R. E. Darrell, Pulse Voltmeter Design and Analysis.
112. PB24921, Navy Dept., PPI repeater VD.
113. PB32740, Rad. Lab., Spectrum analyzer for pulsed oscillators at 3000 Mc.
114. PB32786, Britton Chance, Precision Time Calibrator and Range Measuring System.

Pulse Tubes

121. R. L. Sproull, An Investigation of Short-time Thermionic Emission from Oxide-coated Cathodes, *Phys. Rev.*, Vol. 67, 166 (March 1945).
122. Editor, Cavity Magnetrons, *Electronics*, Vol. 19, No. 1, 126–131 (January 1946).
123. H. G. Shea, Theory of Magnetron Tubes and Their Uses, *Elec. Ind.*, Vol. 5, No. 1, 66–70 (January 1946).
124. J. H. Findlay, Design and Construction of Radar Series Spark-gap Tubes, *Elec. Mfg.*, Vol. 37, No. 3, 131–132 (March 1946).
* 125. J. B. Fisk, H. D. Hagstrum, and P. L. Hartman, The Magnetron as a Generator of Centimeter Waves, *B.S.T.J.*, Vol. 25, 167–348 (April 1946).
126. J. T. Randall, The Cavity Magnetron, *Proc. Phys. Soc.*, May 1, 1946, pp. 247–252.
* 127. R. R. Law, D. G. Burnside, R. P. Stone, and W. B. Whalley, Development of Pulse Triodes to Give 1 Megawatt at 600 Megacycles, *R.C.A. Rev.*, Vol. 7, 253–264 (June 1946).
128. M. Levy, Power Pulse Generator, *Wireless Eng.*, Vol. 23, 192–197 (July 1946).
* 129. Harold Heins, Hydrogen Thyratrons, *Electronics*, Vol. 19, No. 7, 96–102 (July 1946).
130. E. A. Coomes, The Pulsed Properties of Oxide Cathodes, *J. Applied Phys.*, Vol. 17, 647–657 (August 1946).
131. J. J. Glauber, Radar Vacuum-tube Developments, *Elec. Comm.*, Vol. 23, 306–319 (September 1946).
132. Lloyd P. Hunter, Energy Build-up in Magnetrons, *J. Applied Phys.*, Vol. 17, 833–843 (October 1946).
133. PB28647, R. C. Fletcher and G. M. Lee, Preliminary Studies of Magnetron Build-up.

Timing Circuits and Relaxation Oscillators

* 141. H. J. Reich, Trigger Circuits, *Electronics*, Vol. 12, No. 8, 14 (August 1939).
* 142. O. S. Puckle, *Time Bases*. John Wiley (1943).
143. F. E. Terman, *Radio Engineers' Handbook*, pp. 511–516. McGraw-Hill (1943).
* 144. Navy Dept., Timing Circuits, *Navships* 900,013 (May 1944).
145. M. V. Kiebert and A. F. Inglis, Multivibrator Circuits, *Proc. I.R.E.*, Vol. 33, 534–539 (August 1945). (Calculation of period from time constants.)
* 146. D. McMullan, A Single Sweep Time Base, *Electronic Eng.*, Part 1, Vol. 65, No. 1, 21–23 (January 1946); Part 2, Vol. 65, No. 2, 47–50 (February 1946).
147. L. B. Tooley, Gate Circuits for Chronographs, *Electronics*, Vol. 19, No. 5, 144–145 (May 1946).
148. R. Stang, The Blocking Oscillator, *Radio News*, Vol. 36, No. 3, 14–15, 20 (September 1946).

Delay Lines

151. J. P. Blewett and others, *Delay Lines*. General Electric Co. reports (1943).
152. M. J. DiToro, *Delay Lines*. Hazeltine Electronics Corp. reports (1945).
153. M. G. E. Golay, The Ideal Low-pass Filter in the Form of a Dispersionless Lag Line, *Proc. I.R.E.*, Vol. 34, 138P–144P (March 1946).
154. J. M. Lester, Transient Delay Line, *Electronics*, Vol. 19, No. 4, 147–149 (April 1946).
155. H. E. Kallmann, High-impedance Cable, *Proc. I.R.E.*, Vol. 34, 348–351 (June 1946).
156. H. E. Kallmann, Equalized Delay Lines, *Proc. I.R.E.*, Vol. 34, 646–657 (September 1946).
157. K. H. Zimmerman, Spiral Delay Lines, *Elec. Comm.*, Vol. 23, 327–328 (September 1946).

Pulse Transformers

161. H. A. Wheeler, Formulas for the Skin Effect, *Proc. I.R.E.*, Vol. 30, 412–424 (September 1942).
162. R. Lee, Iron-core Components in Pulse Amplifiers, *Electronics*, Vol. 16, No. 8, 115–117, 262–268 (August 1943).
* 163. A. G. Ganz, Permalloy Tape in Wide-band Telephone and Pulse Transformers, *Elec. Eng.*, Vol. 65, *Trans.*, pp. 177–183 (April 1946).
164. PB32773, J. W. Dunifon, Pulse Transformers.

Computers

171. J. T. Potter, A Four-tube Counter Decade, *Electronics*, Vol. 17, No. 6, 110–113, 358, 360 (June 1944).
172. Editor, Preset Interval Timer, *Elec. Ind.*, Vol. 4, No. 7, 97–99, 130, 134, 138, 142, 146 (July 1945).
173. V. H. Regener, Directly Coupled Pentode Trigger Pairs, *Rev. Sci. Instr.*, Vol. 17, 180–184 (May 1946). Decade Counting Circuits, 185–189.
174. A. W. Burke, Super Electronic Computing Machine, *Elec. Ind.*, Vol. 5, No. 7, 62–67, 96 (July 1946).
175. H. G. Shea, Electronic True Decade Counters, *Elec. Ind.*, Vol. 5, No. 9, 82–84, 136 (September 1946).
176. A. Kip, A. Bousquet, R. Evans, and W. Tuttle, Design and Operation of an Improved Counting Rater Meter, *Rev. Sci. Instr.*, Vol. 17, 323–333 (September 1946).
177. B. E. Watt, Current Integrator, *Rev. Sci. Instr.*, Vol. 17, 334–338 (September 1946).
178. A. G. Bousquet, Radioactivity Meter for Nuclear Research, *Elec. Ind.*, Vol. 5, No. 9, 88–89 (September 1946).
179. I. E. Grosdoff, Electronic Counters, *R.C.A. Rev.*, Vol. 7, 438–447 (September 1946).
180. V. H. Regener, Reversible Decade Counting Circuit, *Rev. Sci. Instr.*, Vol. 17, 375–376 (October 1946).
* 181. R. R. Batcher and William Moulic, *The Electronic Control Handbook*. Caldwell-Clements (1946). Section 3, Chapter 3, Counting and Timing Controls, pp. 178–195, 338.
182. S. A. Karff, *Electron and Nuclear Counters*. Van Nostrand (1946).

Picture Transmission

* 191. RCA, *Television*, Vol. 1 (July 1936); Vol. 2 (October 1937).
* 192. J. C. Wilson, *Television Engineering*. Sir Isaac Pitman & Sons, London (1937). Chapter 12, Physical Limitations, pp. 420–437.

193. J. C. Wilson, Channel Width and Resolving Power in Television Systems, *J. Television Soc.*, Vol. 2, 397–420 (June 1936).
* 194. RCA, *Radio Facsimile*, Vol. 1 (October 1938).
* 195. V. K. Zworykin and G. A. Morton, *Television*. John Wiley (1940).
* 196. National Television Systems Committee, *Television Standards and Practice* (1943).
* 197. C. E. Dean, *Television Principles*. Hazeltine Electronics Corp. (1944).

Pulse Modulation for Communication

201. R. A. Heising, Transmission System, U. S. Pat. 1,655,543, April 18, 1924–Jan. 10, 1928.
202. R. H. Ranger, Reproducing and Transmitting Pictures, U. S. Pat. 1,848,839, Feb. 26, 1924–March 8, 1932.
203. J. L. Finch, Signaling System, U. S. Pat. 1,887,237, May 31, 1929–Nov. 8, 1932.
204. R. D. Kell, Signaling System, U. S. Pat. 2,061,734, Sept. 29, 1934–Nov. 24, 1936.
205. W. R. Koch, Secret Communication System, U. S. Pat. 2,199,634, June 21, 1938–May 7, 1940.
206. D. G. C. Luck, Signaling System, U. S. Pat. 2,227,596, March 31, 1938–Jan. 7, 1941.
207. Bertram Trevor, Pulse Signaling System, U. S. Pat. 2,361,437, Dec. 4, 1940–Oct. 31, 1944.
* 208. E. M. Deloraine and E. Labin, Pulse-time Modulation, *Electronics*, Vol. 18, No. 1, 100–104 (January 1945).
* 209. E. M. Deloraine and E. Labin, Pulse-time Modulation, *Elec. Com.*, Vol. 22, 91–98 (1944).
210. Editor, Pulse Position Modulation Technique, *Elec. Ind.*, Vol. 4, No. 12, 82–87, 180, 182, 184, 186, 188, 190 (December 1945).
211. D. D. Grieg, Multiplex Broadcasting, *Elec. Com.*, Vol. 23, 19–26 (March 1946).
212. F. F. Roberts and J. C. Simmons, correspondence, *Wireless Eng.*, Vol. 23, 93 (March 1946).
* 213. J. J. Kelleher, Pulse-modulated Radio Relay Equipment, *Electronics*, Vol. 19, No. 5, 124–129 (May 1946).
* 214. D. D. Grieg and A. M. Levine, Pulse-time-modulated Multiplex Radio Relay System-terminal Equipment, *Elec. Com.*, Vol. 23, 159–178 (June 1946).
215. Editor, Multi-channel Pulse Modulation, *Wireless World*, Vol. 52, 187–192 (June 1946).
216. D. I. Lawson, A. V. Lord, and S. R. Kharbanda, Transmitting Sound on the Vision Carrier of a Television System, *J. I.E.E.*, Part III, Vol. 93, 251–274 (July 1946).
217. W. R. Greer, Pulse Modulating System, *Electronics*, Vol. 19, No. 9, 126–131 (September 1946).
218. D. G. Tucker, Pulse Distortion: Interchannel Interference in Multichannel Systems, *J. I.E.E.*, Vol. 93, 323–334 (September 1946).
219. H. S. Black, J. W. Beyer, T. J. Grieser, F. A. Polkinghorn, A Multichannel Microwave Radio Relay System, *Elec. Eng.*, Vol. 65, 798–805 (December 1946).
* 220. RCA, *Television*, Vol. 3 (December 1946); Vol. 4 (January 1947).
221. Airborne Radar Specifications, *Electronics*, Vol. 20, 132 (February 1947).
222. D. G. Fink, *Principles of Television Engineering*, McGraw-Hill (1947).
* 223. D. G. Fink, *Radar Engineering*, McGraw-Hill (1947).
224. Harvard Univ., *Electronic Tubes and Circuits*, McGraw-Hill (1947).
225. Reuben Lee, *Electronic Transformers and Circuits*, John Wiley (1947).
226. W. M. Goodall, Telephony by Pulse Code Modulation, *Bell Sys. Tech. J.*, Vol. 26, 395–409 (July 1947).
227. R. W. Sears, Electron Beam Deflection Tube for Pulse Code Modulation, *Bell Sys. Tech. J.*, Vol. 27, 44–57 (January 1948).
228. L. B. Arguimbau, *Vacuum-tube Circuits*, John Wiley (1948).
229. Australia, *Textbook of Radar*, Chapman & Hall (1948).
230. S. Goldman, *Frequency Analysis, Modulation and Noise*, McGraw-Hill (1948).
231. E. C. Pollard and J. M. Sturtevant, *Microwaves and Radar Electronics*, John Wiley (1948).
232. R. Q. Smith, *Radio Aids to Navigation*, Macmillan (1948).
233. Denis Taylor and C. H. Westcott, *Principles of Radar*, Macmillan (1948).
234. C. E. Shannon, A Mathematical Theory of Communication, *Bell Sys. Tech. J.*, Vol. 27, 379–423, 623–656 (July, October 1948).
235. B. M. Oliver, J. R. Pierce, and C. E. Shannon, The Philosophy of Pulse-code Modulation, *Proc. I.R.E.*, Vol. 36, 1324–1331 (November 1948).
236. C. E. Shannon, Communication in the Presence of Noise, *Proc. I.R.E.*, Vol. 37, 10–21 (January 1949).
237. H. Goldberg and C. C. Bath, Multiplex Employing Pulse-time and Pulsed-frequency Modulation, *Proc. I.R.E.*, Vol. 37, 22–28 (January 1949).
238. A. W. Friend, Theory and Practice of Tropospheric Sounding by Radar, *Proc. I.R.E.*, Vol. 37, 116–138 (February 1949).
239. J. H. Dewitt, Jr., and E. K. Stodola, Detection of Radio Signals Reflected from the Moon, *Proc. I.R.E.*, Vol. 37, 229–242 (March 1949).
240. W. G. Tuller, Theoretical Limitations on the Rate of Transmission of Information, *Proc. I.R.E.*, Vol. 37, 468–478 (May 1949).
241. E. M. Deloraine, Pulse Modulation, *Proc. I.R.E.*, Vol. 37, 702–705 (June 1949).
* 242. *M.I.T. Radiation Laboratory Series*. 28 volumes. McGraw-Hill.

SECTION 10

TRANSMISSION CIRCUITS

TRANSMISSION CIRCUITS

WIRE TRANSMISSION LINES

By John D. Taylor

1. TYPES OF COMMUNICATION TRANSMISSION CIRCUITS

Wire communication circuits are classified in two groups, open (bare) wire and cable. Each group includes a number of different gages and types of wire, and the electrical characteristics of each gage and type are different. Both groups have their respective fields of use, and frequently these fields overlap and supplement each other. Open wire is generally economical where the circuit requirements are relatively small and cable costs would be prohibitive. Cable is desirable and economical where large groups of circuits and the higher-frequency types of services are involved, and where maximum protection from surface interferences is essential.

Cable is also employed for specific applications; for example, coaxial cable for television and radio transmission lines.

2. FREQUENCY SPECTRUM

Communication intelligence and signals (with the exception of those d-c signals used in telegraph, emergency, control, and other services) are usually transmitted between points in the form of a-c electrical energy or wave propagation of definite frequencies or frequency ranges. Because of the different devices for and methods of generating and transmitting these various a-c frequencies, and in order to avoid interference between the numerous circuits and services, frequency allocations have been determined for the various classes of communication and signal facilities.

Communication services may be classified, for the purpose of assigning suitable facilities, according to the frequency range within which they operate. Figure 1 is a chart of frequencies employed in power and communication services up to about 5 mc. (See Sect. 17, art. 8.)

3. ELECTRICAL CHARACTERISTICS

Primary constants of wire lines consist of resistance r in ohms, inductance L in henries, capacitance C in farads, and leakage conductance g in mhos. For convenience, L, C, or g may be expressed in smaller units, such as microfarads (millionths of a farad) for C. These constants vary differently for different wire materials, sizes, spacings, insulation, conductivity, temperature, and moisture conditions, as well as for different types of construction, as indicated in Table 1.

Secondary constants are values derived from primary constants; for communication purposes, they consist principally of (1) impedance z in ohms and phase angle, (2) propagation constant, composed of the attenuation and wave length constants, (3) wave length, (4) velocity of propagation, and (5) cut-off frequency. From the various constants, the electrical characteristics of the different wire lines have been determined, and these characteristics are useful in considering the suitability of such lines for communication purposes and in the engineering of the wire communication plant.

The secondary constants, expressed in terms of the primary constants, are as indicated in the following equations. These equations apply to uniform lines of infinite length or terminated in their characteristic impedances.

CHARACTERISTIC IMPEDANCE, $Z_0 \angle \theta$. A wire transmission line has both series impedance, z, and shunt admittance, y. For a line of uniformly distributed primary constants and of infinite length $z = r + j\omega L$ and $y = g + j\omega C$, where ω (angular velocity) $= 2\pi f$ radians [1 radian $= (360/2\pi)°$].

$$Z_0 = \sqrt{\frac{z}{y}} = \sqrt[4]{\frac{r^2 + \omega^2 L^2}{g^2 + \omega^2 C^2}} \quad \text{ohms} \tag{1}$$

and

$$\angle \theta = \frac{1}{2}\left(\tan^{-1}\frac{\omega L}{r} - \tan^{-1}\frac{\omega C}{g}\right)$$

Table 1. Characteristics of Representative Types of Wire Telephone Lines at 1000 Cycles (1)

Type of Line	Gage of Wire, mils	Spacing of Wires, in.	Type of Loading *	Distributed (2) Constants per Loop Mile				Propagation Constant				Line Impedance				Wave-length, miles	Velocity, thousands of miles per second	(5) Cut-off Frequency, cycles	Attenuation, db per mile
				r, ohms (68° F for o.w., 55° F for ca.)	L, henries	(3) C μf	(4) g μmhos	Polar: Magnitude	Polar: Angle, degrees +	Rectangular: α nepers	Rectangular: β radians	Polar: Magnitude	Polar: Angle, degrees −	Rectangular: r, ohms	Rectangular: x, ohms −				
Open Wire (non-pole pair, hard-drawn copper)																			
Physical	165	8		4.11	0.00311	0.00996	0.14	0.0353	83.99	0.00370	0.0351	565	5.88	562	58	179.0	179.0		0.032
Side	165	12	Open	4.11	0.00337	0.00915	0.29	0.0352	84.36	0.00346	0.0350	612	5.35	610	57	179.5	179.5		0.030
Phantom	165	12	wire	2.06	0.00208	0.01514	0.58	0.0355	85.34	0.00288	0.0354	373	4.30	372	28	177.5	177.5		0.025
Side	128	12	is	6.74	0.00353	0.00871	0.29	0.0356	81.39	0.00533	0.0352	650	8.32	643	94	178.5	178.5		0.046
Phantom	128	12	not	3.37	0.00216	0.01454	0.58	0.0357	82.84	0.00445	0.0355	401	6.73	398	47	177.0	177.0		0.039
Side	104	12	loaded	10.15	0.00366	0.00837	0.29	0.0363	77.93	0.00760	0.0355	692	11.75	677	141	177.0	177.0		0.066
Phantom	104	12		5.08	0.00223	0.01409	0.58	0.0363	79.84	0.00640	0.0357	421	9.70	415	71	176.0	176.0		0.056
Cable (quadded toll)																			
Side	19	......	NL	85.8	0.001	0.062	1.5	0.1830	46.98	0.1249	0.1338	470	42.80	345	319	46.9	46.9	......	1.085
Side	16	......	NL	42.1	0.001	0.062	1.5	0.1288	49.13	0.0842	0.0974	331	40.65	251	215	64.5	64.5	......	0.730
Side	19	......	H-44-S	89.4	0.039	0.062	1.5	0.3188	79.87	0.0561	0.3138	818	9.91	806	141	20.0	20.0	5705	0.487
Phantom	19	......	H-25-P	44.7	0.023	0.100	2.4	0.3082	81.30	0.0466	0.3047	491	8.48	485	72	20.6	20.6	5916	0.405
Side	19	......	H-88-S	92.2	0.078	0.062	1.5	0.4408	84.56	0.0418	0.4388	1131	5.22	1126	103	14.3	14.3	3997	0.363
Phantom	19	......	H-50-P	46.2	0.045	0.100	2.4	0.4243	85.25	0.0351	0.4228	675	4.53	673	53	14.9	14.9	4193	0.305
Side	16	......	H-44-S	45.7	0.039	0.062	1.5	0.3148	84.61	0.0296	0.3134	808	5.17	805	73	20.0	20.0	5705	0.257
Phantom	16	......	H-25-P	22.8	0.023	0.100	2.4	0.3032	85.41	0.0243	0.3022	483	4.37	481	37	20.8	20.8	5916	0.211
Side	16	......	H-88-S	48.5	0.078	0.062	1.5	0.4380	87.64	0.0224	0.4374	1124	2.71	1123	53	14.4	14.4	3997	0.194
Phantom	16	......	H-50-P	24.3	0.045	0.100	2.4	0.4223	87.43	0.0189	0.4218	672	2.35	672	28	14.9	14.9	4193	0.164

* The type of loading is designated as follows: First a letter indicating the spacing of coils: H = 6000 feet; next the number of millihenries per coil; then S or P for side or phantom. See Section 17, art. 16.

Notes: (1) All values are dry weather, calculated and approximate. Values actually vary from cable to cable and pair to pair in the same cable.
(2) Assumed distributed for loaded cable.
(3) Assumes 40-wire o.w. line.
(4) DP in. for 12-in. and CS in. for 8-in. spacing.
(5) Exceeds open-wire system requirements.

For standard non-loaded telephone cable lines, L and g are relatively small, so that for approximate computations these constants may be neglected and

$$Z_0 \doteq \sqrt{\frac{r}{\omega C}} \angle 45^\circ \quad \text{ohms} \tag{2}$$

For lumped loaded cable lines (the usual method of loading such lines), the midsection characteristic impedance Z_0 of an infinite loaded line may be determined from a pi network, converted to a T network, which is electrically equivalent to the combined lumped and distributed constants of a complete loading section of the line. The a (series) and

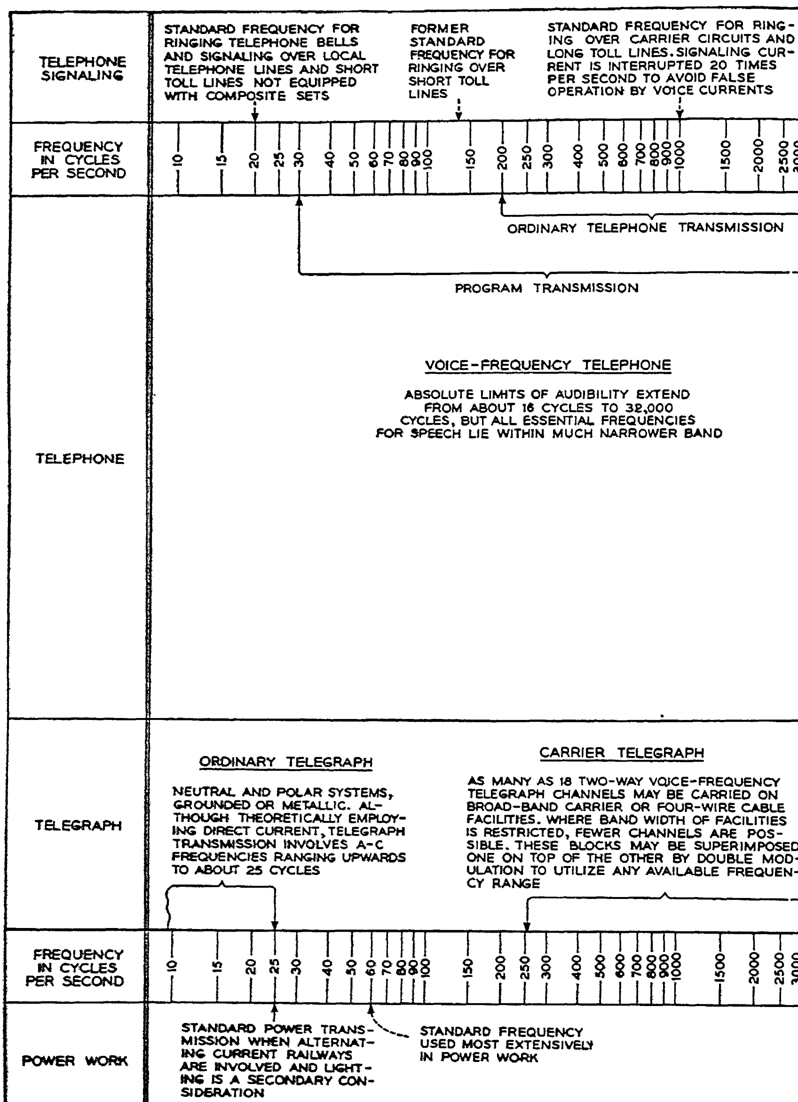


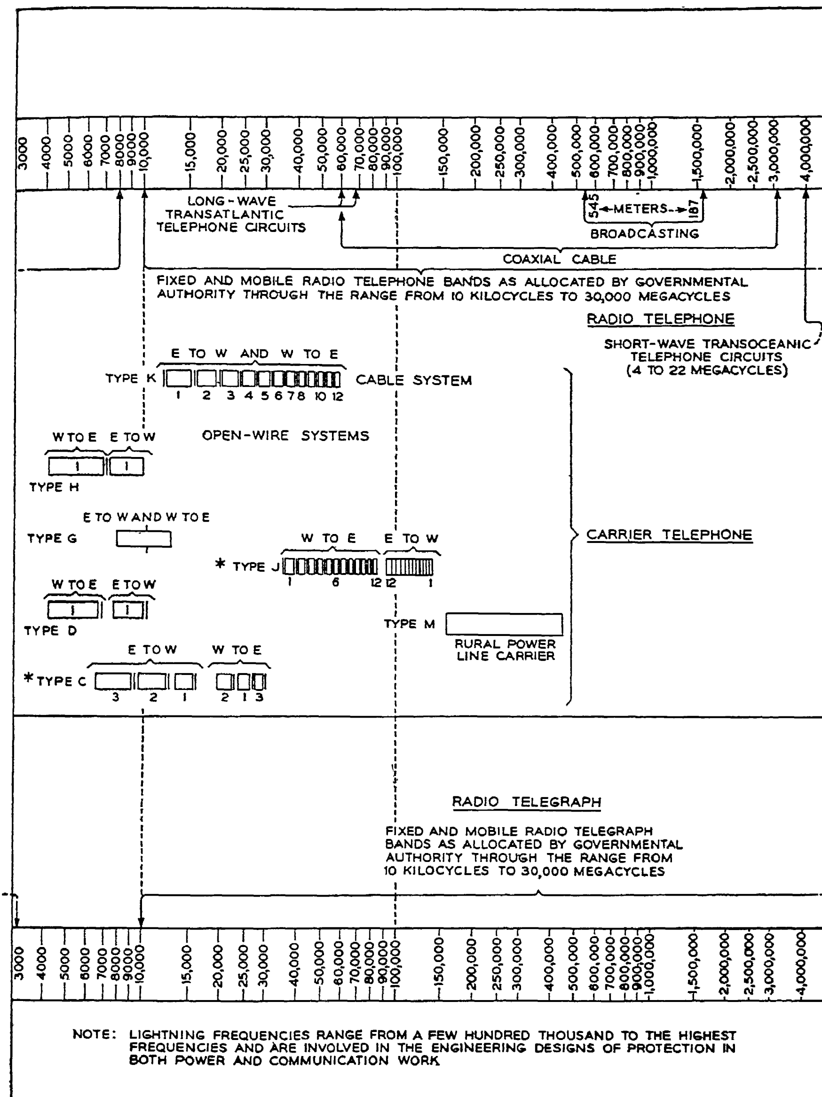
FIG. 1. Operating Frequencies for Power and

b (shunt) values of the T network having thus been determined,

$$Z_0 = \sqrt{ab + \frac{a^2}{4}} \quad \text{ohms} \tag{3}$$

For uniform or lumped loaded cable lines at voice frequencies ωL is large with respect to r and

$$Z_0 \doteq \sqrt{\frac{L}{C}} \angle 0° \quad \text{ohms} \tag{4}$$



* THESE REPRESENT THE FREQUENCY ALLOCATIONS FOR ONE SYSTEM OF THIS TYPE; OTHERS ARE DISPLACED SLIGHTLY, OR EMPLOY OTHER SIDEBAND

Communication Services (Courtesy Bell System)

For non-loaded open wire lines at radio frequencies

$$Z_0 \doteq 276 \log_{10} \frac{2D}{d} \quad \text{ohms} \tag{5}$$

where D is the distance between the centers of the wires and d is the diameter of the wires.

For coaxial (concentric tube) cable lines at carrier and higher frequencies

$$Z_0 \doteq 138 \log_{10} \frac{b}{a} \quad \text{ohms} \tag{6}$$

where a is the outer radius of the inner conductor and b is the inner radius of the outer conductor.

PROPAGATION CONSTANT. The propagation constant, γ, is a function of the series impedance, z, and the admittance, y, in the vector relation

$$\gamma = \sqrt{zy} = \sqrt{(r + j\omega L)(g + j\omega C)} \tag{7}$$

Also, γ is composed of an attenuation constant α and a wave length constant β in the relation

$$\gamma = \alpha + j\beta = \gamma \cos\theta + j\gamma \sin\theta \tag{8}$$

where

$$\alpha = \sqrt{1/2\sqrt{(r^2 + \omega^2L^2)(g^2 + \omega^2C^2)} + 1/2\,(gr - \omega^2LC)} \tag{8a}$$

$$\beta = \sqrt{1/2\sqrt{(r^2 + \omega^2L^2)(g^2 + \omega^2C^2)} - 1/2\,(gr - \omega^2LC)} \tag{8b}$$

Equation (8a) gives α in nepers (1 neper = 8.686 db), and eq. (8b) gives β in radians (1 radian = 57.296°). It is usually more convenient to calculate α and β from eqs. (7) and (8) than from eqs. (8a) and (8b).

From the above discussion it is seen that the propagation constant represents both the dying-away and phase-change effects of the voltages and currents, as they progress along the line.

If γ is given for 1 mile, then the total attenuation and phase change effects for l miles is $l\gamma$. The ratio of the current I_2 at the receiving end to the current I_1 at the sending end of a uniform line, l miles long and properly terminated, is

$$\frac{I_2}{I_1} = \varepsilon^{-l\gamma} = \varepsilon^{-l(\alpha+j\beta)} = \varepsilon^{-l\alpha}/\overline{l\beta} \tag{9}$$

Thus, the magnitude of $I_2 = I_1\varepsilon^{-l\alpha}$ and the two currents differ in phase by the angle $l\beta$, which is expressed in radians.

From Eq. (9)

$$2.303 \log_{10} \frac{I_2}{I_1} = -l\alpha$$

and the magnitude of the propagated current I_2, in terms of the sending current I_1, at any point along a uniform transmission line, is

$$\frac{I_2}{I_1} = \frac{1}{\log_{10}^{-1} \dfrac{l\alpha}{2.303}} \tag{10}$$

Also, if P_2 is the received power and E_1 and I_1 are the sending end voltage and current, respectively, with phase angle θ, then

$$P_2 = E_1 I_1 \varepsilon^{-2l\alpha} \cos\theta \tag{11}$$

showing that the power is attenuated in accordance with the square of the current ratio.

If the constants of the transmission line are such that $r/L = g/C$ or $rC = Lg$, then the line attenuation and velocity of propagation do not change appreciably with frequency and the line is said to be *distortionless*, since

$$\alpha \doteq \sqrt{rg} \text{ and } W \doteq \frac{1}{\sqrt{LC}} \tag{12}$$

For standard non-loaded telephone cables, L and g are considered negligible; hence from eqs. (8)

$$\alpha \doteq \beta \doteq \sqrt{\frac{\omega r C}{2}} \quad \text{nepers (for } \alpha) \quad \text{and} \quad \text{radians (for } \beta) \tag{13}$$

For uniformly or lumped loaded cable lines (up to about mid-frequency range) where r and g are small compared to ωL and ωC respectively, an approximate expression for α is

$$\alpha \doteq \frac{r}{2}\sqrt{\frac{C}{L}} + \frac{g}{2}\sqrt{\frac{L}{C}} \doteq \sqrt{CL}\left(\frac{r}{2L} + \frac{g}{2C}\right) \quad \text{nepers} \tag{14}$$

In this equation, $r/2L$ and $g/2C$ are the damping constants of the series and shunt constants, respectively, of the line.

If g is assumed to be zero, eq. (14) becomes

$$\alpha \doteq \frac{r}{2}\sqrt{\frac{C}{L}} \doteq \sqrt{\frac{r}{2L} \cdot \frac{rC}{2}} \quad \text{nepers} \tag{15}$$

Equation (15) also applies for open-wire lines in the r-f range.

Thus, it is seen from eqs. (13) and (15) that, for values of $(r/2L) < \omega$, α is decreased by increasing L, i.e., by loading the cable line, although the full advantage is not realized, because, in loading the line, r is increased slightly by the resistance of the loading coils.

For coaxial (concentric tube) cable lines in the r-f range

$$\alpha \doteq \frac{r}{2Z_0} \quad \text{nepers} \tag{16}$$

and, where both conductors are of the same material and the line dimensions are available,

$$\alpha \doteq \frac{\sqrt{\rho\mu f}(1/a + 1/b)}{276 \log_{10}(b/a)} \cdot 10^{-9} \quad \text{neper per unit length} \tag{17}$$

where ρ = resistivity of conductors in emu (about 1730 emu for pure copper).
μ = magnetic permeability of the insulation.
f = frequency in cycles per second.
a = outer radius of the inner conductor.
b = inner radius of the outer conductor.

WAVE LENGTH. The phase of the voltage and current for a uniform line is continually changing in a modified sine-wave pattern, as progression takes place along the line. A complete phase change of 2π radians (360°) will occur in the length of line traversed by the voltage and current during the time they pass through 1 cycle. Thus, the wavelength λ for a particular line may be equated as

$$\lambda = \frac{2\pi}{\beta} \quad \text{miles} \tag{18}$$

β being the phase change in radians per mile.

VELOCITY OF PROPAGATION. Since a wave length is the length traversed by the voltage and current during 1 cycle, the velocity of propagation W may be equated as

$$W = \lambda f \quad \text{miles per second} \tag{19}$$

f being the frequency in cycles per second.

For loaded cable lines, the inductance and capacitance of the line have a direct bearing on the wave length constant, wave length, and velocity of propagation. Since r and g are usually negligible for such lines, eq. (8b) for β becomes

$$\beta \doteq \omega\sqrt{LC} \doteq 2\pi f\sqrt{LC} \tag{20}$$

Since $\lambda = 2\pi/\beta$,

$$W = \lambda f = \frac{2\pi f}{\beta} \doteq \frac{1}{\sqrt{LC}} \quad \text{miles per second} \tag{21}$$

Note: If L and C are expressed as the values for one load section, then β and λ are in loads and W is in loads per second.

CUTOFF FREQUENCY. For non-loaded lines (cable or open wire), the cutoff frequency is usually high enough for all practical purposes for the service for which these facilities will be used, but for loaded cable lines the cutoff frequency is an important factor and may be expressed as

$$f_c = \frac{1}{\pi\sqrt{LC}} \tag{22}$$

where L is the inductance of the loading coil in henries and C is the capacitance of the loading section in farads for any particular line.

Thus, the periodic lumped loaded line transmits all frequencies up to a critical or cutoff frequency. However, in the actual line there are some deviations from the ideal loading and some resistance in the line and loading coils, resulting in the attenuation of the transmitted frequency range to some degree.

4. EQUIVALENT NETWORKS

Equivalent networks for a length of uniform line, transmitting alternating currents, may be constructed in the T or pi (π) form.

For the T network, a, the value of each of the two series arms, and b, the value of the shunt arm, may be obtained from the equations

$$a = Z_0 \tanh \frac{l\gamma}{2} \tag{23a}$$

$$b = \frac{Z_0}{\sinh l\gamma} \tag{23b}$$

where Z_0, l, and γ are the characteristic impedance, length of line in miles, and the propagation constant per mile.

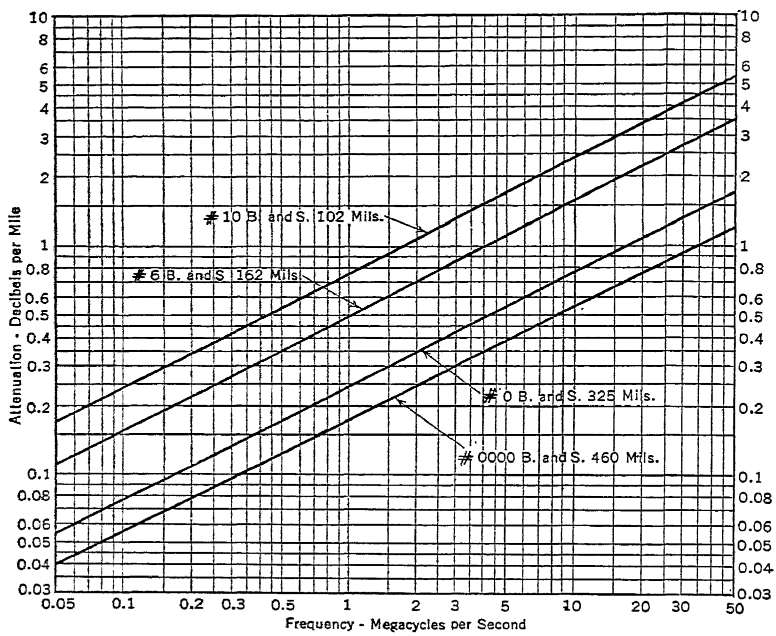


FIG. 2. Calculated Attenuation vs. Frequency Characteristics, 600 ohm, Two-wire Lines, Solid Copper Conductors; Leakage and Radiation Neglected

For the pi network, r, the value of each of the two shunt arms, and s, the value of the series arm, may be obtained from the equations

$$r = Z_0 \coth \frac{l\gamma}{2} \tag{24a}$$

$$s = Z_0 \sinh l\gamma \tag{24b}$$

where Z_0, l, and γ have the same meaning as for the T network.

Figure 2 shows the calculated attenuation-frequency characteristics of 600-ohm, two-wire, solid copper lines, in which leakage conductance and radiation are assumed to be negligible. Calculations are based on the equation

$$\alpha \doteq 8.686 \frac{\sqrt{1830f}}{aZ_0} \cdot 10^{-6} \quad \text{decibel per mile} \tag{25}$$

where a = radii of the various conductors in miles, f = megacycles per second, and Z_0 = characteristic impedances of the various lines. The effect on current distribution of the return wire (important only when the ratio of separation to wire diameter is less than 20 to 1) is neglected in the equation.

Figure 3 shows the calculated attenuation-frequency characteristics of copper concentric tube lines having optimum ratios of conductor radii. Leakage is neglected. Calculations are based on eq. (17). Assuming both conductors to be of the same material, α will be

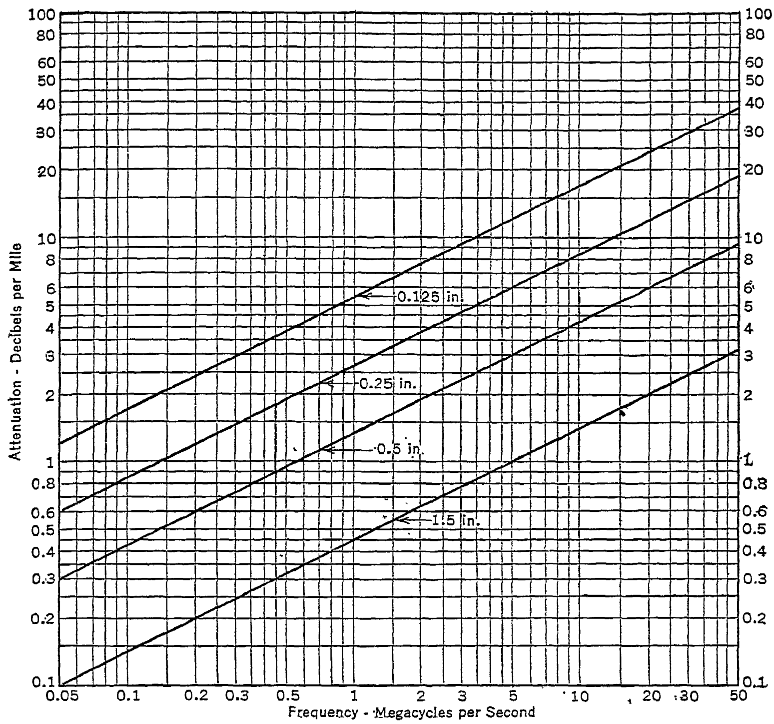


FIG. 3. Calculated Attenuation vs. Frequency Characteristics, Copper Concentric Tube Lines. Optimum ratio of conductor radii. Leakage neglected. Dimension indicated is inner radius of outer conductor.

a minimum when the ratio $b/a = 3.6$. With $\rho = 1730$ and α a minimum, eq. (17) can be reduced to

$$\alpha \doteq \frac{0.686\sqrt{f}}{b} \quad \text{decibels per mile} \tag{26}$$

where f is in megacycles per second and b is the inner radius of the outer conductor in inches.

WAVE GUIDES—THEORY

By S. A. Schelkunoff

Definition. A wave guide is a structure consisting of either conductors or dielectrics, or both, in which the boundaries between different media are cylindrical (surfaces made up by a translation of a straight line; circular cross-section is not essential). A *metal wave guide* is a wave guide containing at least one conductor. A *dielectric wave guide* is a wave guide consisting of dielectrics only.

5. MODES OF TRANSMISSION

Three types of waves are possible in straight metal wave guides filled with a homogeneous dielectric: transverse electromagnetic or TEM waves, transverse electric or TE waves, and transverse magnetic or TM waves. In TEM waves the electric and magnetic vectors are perpendicular to the direction of the guide; in TE waves the electric vector is so disposed; and in TM waves the magnetic vector is perpendicular to the guide. TEM waves are possible only if the guide consists of two or more separate conductors; there are no restrictions on the existence of the other types. If the dielectric is non-homogeneous, then in general the waves are of *hybrid type*, with all components of E and H vectors present.

Each wave guide permits an infinite number of transmission modes. Besides the abovementioned general characteristics, each mode is distinguished by a transverse field pattern consistent with the structure of the guide. Theoretically, at least, it is possible to excite an arbitrary field pattern over a given cross-section of the guide; but, in general, this pattern will not be maintained along the guide. The self-maintaining patterns are the ones that define the various modes of transmission. In general, these patterns depend on the frequency as well as the structure of the wave guide; but, if the dielectric is homogeneous, then the patterns are independent of the frequency and are given solely by the geometry of the metal boundaries. Each self-maintaining pattern is either attenuated or is traveling with a phase velocity peculiar to it. (See Propagation Constants, article 6.) An arbitrary pattern excited over a given cross-section breaks up into self-maintaining patterns. The field at various distances from the source is the result of interference between the self-maintaining patterns arriving with different amplitudes and phases.

In wave guides with rectangular, circular, and elliptic boundaries, the various TE and TM modes are designated by a double subscript, TE_{mn} and TM_{mn}, where m and n are integers appearing in the mathematical functions describing transverse field patterns. For each shape of the guide, indices m and n reflect certain physical characteristics of the wave; but the same indices for different shapes correspond to waves with different characteristics, and waves with similar characteristics may have different indices. This happens because a gradual deformation of a circular boundary into a rectangular affects different modes differently.

In a wave guide of arbitrary shape the various modes are designated by a single subscript which denotes the position of the "cutoff frequency" on the frequency scale. The *dominant wave* is the wave with the lowest cutoff frequency.

6. PROPAGATION CONSTANTS OF IDEAL WAVE GUIDES

For an ideal non-dissipative wave guide with a homogeneous dielectric the propagation constant γ_g along the guide for a given transmission mode is

$$\gamma_g = \omega_c\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{\omega}{\omega_c}\right)^2} \qquad \omega = 2\pi f \tag{1}$$

where the *cutoff frequency* f_c is determined by the permeability μ and dielectric constant ε and by the geometry of the boundaries. For $f < f_c$, γ is real and the wave is attenuated even though there is no dissipation of energy. For $f > f_c$ the above equation becomes

$$\gamma_g = i\omega\sqrt{\mu\varepsilon}\sqrt{1 - \left(\frac{\omega_c}{\omega}\right)^2}$$

and the propagation constant is imaginary. The wave becomes *active* in transmitting energy to large distances since under the assumed ideal conditions the amplitude of the wave does not diminish. These characteristics are the characteristics of a high-pass filter.

For wave guides with a non-homogeneous dielectric there is no simple general expression for the propagation constant.

PHASE VELOCITY. Above the cutoff the phase velocity along the guide is

$$v_g = \frac{v}{\sqrt{1 - (\omega_c/\omega)^2}} \qquad v = \frac{1}{\sqrt{\mu\varepsilon}} = \frac{c}{\sqrt{\varepsilon_r}} \tag{2}$$

where v is the intrinsic velocity of the dielectric (velocity of light in the dielectric), c is the intrinsic velocity of vacuum, and ε_r is the dielectric constant relative to vacuum. In hollow wave guides the phase velocity is always higher than the velocity of light in free space; this is in keeping with the high-pass characteristics of wave guides.

GROUP VELOCITY. The group velocity or the velocity of a "wave packet" is

$$v_{\text{group}} = v\sqrt{1 - \left(\frac{\omega_c}{\omega}\right)^2} \tag{3}$$

The product of the group and phase velocities equals the square of the intrinsic velocity.

WAVELENGTH IN THE GUIDE. The wavelength λ_g in the guide (the distance from crest to crest of the wave) is

$$\lambda_g = \frac{v_g}{f} = \frac{(v/f)}{\sqrt{1-(f_c/f)^2}} = \frac{\lambda_v}{\sqrt{\varepsilon_r[1-(f_c/f)]^2}} \tag{4}$$

where λ_v is the free-space wavelength corresponding to the given frequency f and ε_r is the dielectric constant relative to vacuum.

7. RECTANGULAR WAVE GUIDES

FIELDS. In a metal wave guide the effect of the conductivity of the walls on the field distribution is negligible. The following expressions are for perfectly conducting walls, assuming that the coordinate system is disposed as in Fig. 1. The time factor $\exp j\omega t$ is omitted.

$\mathrm{TE}_{m,n}$ wave (if traveling in the positive z direction):

$$E_x = A\frac{n\pi}{b}\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-\gamma z} \qquad E_y = -A\frac{m\pi}{a}\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-\gamma z}$$

$$H_x = -\frac{E_y}{K_z^{TE}} \qquad H_y = \frac{E_x}{K_z^{TE}} \qquad H_z = A\frac{\chi_{mn}^2}{j\omega\mu}\cos\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-\gamma z} \tag{5}$$

$$K_z = \frac{j\omega\mu}{\gamma} \qquad \gamma = \sqrt{j\omega\mu(g+j\omega\varepsilon)+\chi_{mn}^2} \qquad \chi_{mn} = \sqrt{\frac{m^2\pi^2}{a^2}+\frac{n^2\pi^2}{b^2}}$$

where m, n are integers, not equal to zero simultaneously. The quantities g, μ, and ε are respectively the conductivity, the permeability, and the dielectric constant of the dielectric in the guide. Figure 2 shows electric lines for some TE waves; they are also the lines of constant H_z. The transverse H component is perpendicular to the E vector. The density of lines is proportional to the transverse field components. The pattern of the $\mathrm{TE}_{1,0}$ wave in Fig. 2(a) is the building block for patterns of the $\mathrm{TE}_{m,0}$ waves as illustrated in Fig. 2(c); m is the number of such blocks in the pattern of the $\mathrm{TE}_{m,0}$ wave. In the $\mathrm{TE}_{0,n}$ wave the lines are parallel to the horizontal faces of the guide. The pattern of the $\mathrm{TE}_{1,1}$ wave in Fig. 2(c) is the building block for the pattern of the $\mathrm{TE}_{m,n}$ wave when m and n are different from zero; m is the number of such blocks in the x direction (the horizontal direction), and n is the number of blocks in the y direction.

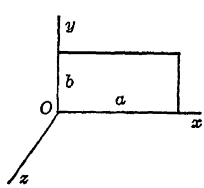


FIG. 1

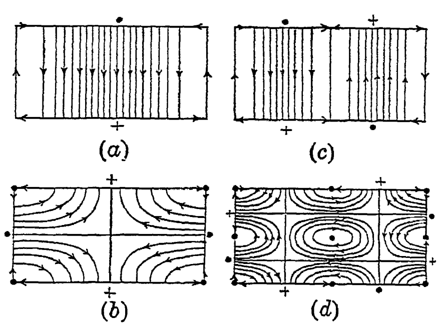


FIG. 2. Electric Lines in Rectangular Guides; (a) $\mathrm{TE}_{1,0}$ Wave; (b) $\mathrm{TE}_{1,1}$ Wave; (c) $\mathrm{TE}_{2,0}$ Wave; (d) $\mathrm{TE}_{2,2}$ Wave

$\mathrm{TM}_{m,n}$ wave (if traveling in the positive z direction):

$$H_x = A\frac{n\pi}{b}\sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-\gamma z} \qquad H_y = -A\frac{m\pi}{a}\cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-\gamma z}$$

$$E_x = K_z^{TM}H_y \qquad E_y = -K_z^{TM}H_x \qquad E_z = A\frac{\chi_{mn}^2}{g+j\omega\varepsilon}\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-\gamma z} \tag{6}$$

$$K_z^{TM} = \frac{\gamma}{g+j\omega\varepsilon} \qquad \gamma = \sqrt{j\omega\mu(g+j\omega\varepsilon)+\chi_{mn}^2} \qquad \chi_{mn} = \sqrt{\frac{m^2\pi^2}{a^2}+\frac{n^2\pi^2}{b^2}}$$

where m, n are integers not equal to zero. Figure 3 shows magnetic lines for some TM waves; they are also the lines of constant E_z. The transverse E component is perpendicular to the H vector. The density of lines is proportional to the transverse field components.

The pattern of the $TM_{1,1}$ wave in Fig. 3(*a*) is the building block for the pattern of the $TM_{m,n}$ wave; m is the number of such blocks in the x direction, and n is the corresponding number in the y direction.

CUTOFF FREQUENCIES. The cutoff frequencies for the various modes of transmission are the frequencies at which the propagation constant vanishes; at these frequencies the propagation constant changes from a real to an imaginary value; these frequencies separate the pass band of frequencies from the stop band. In this sense the cutoff frequencies are defined only for non-dissipative wave guides; the cutoff frequencies of slightly dissipative wave guides are usually defined by neglecting dissipation. In general, the cutoff frequency may be defined as the frequency at which the phase of the propagation constant is 45°; but when the dissipation is really large the concept of "cutoff" loses its practical value.

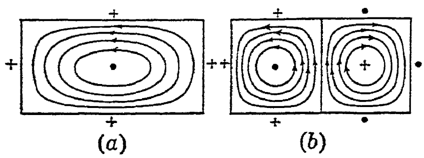


FIG. 3. Magnetic Lines in Rectangular Guides; (*a*) $TM_{1,1}$ Wave; (*b*) $TM_{2,1}$ Wave

For rectangular wave guides the cutoff frequencies and corresponding wavelengths (in vacuum) are

$$f_{mn} = \frac{c}{2\sqrt{\varepsilon_r}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2} \qquad \lambda_{mn} = \frac{2\sqrt{\varepsilon_r}}{\sqrt{(m/a)^2 + (n/b)^2}} \tag{7}$$

where c is the velocity of light in vacuum.

The same formulas hold of TE and TM waves. For the dominant wave

$$\lambda_{1,0} = 2a\sqrt{\varepsilon_r} \qquad f_{1,0} = \frac{c}{2a\sqrt{\varepsilon_r}} \tag{8}$$

where a is the longer dimension of the cross-section of the guide.

ATTENUATION IN THE PASS BAND. In a previous section the exact expression is given for the propagation constant γ when the walls of the guide are perfect conductors. To allow for the imperfect conductivity of the walls the following correction term should be added to γ

$$\Delta\gamma = \alpha_m(1 + j) \tag{9}$$

where α_m is the attenuation in nepers per meter due to absorption of energy by the walls. For TE_{mn} wave

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{60\pi b}\left[\frac{p(pm^2 + n^2)}{p^2m^2 + n^2}\sqrt{1 - \nu_{mn}^2} + \frac{(1 + p)\nu_{mn}^2}{\sqrt{1 - \nu_{mn}^2}}\right] \qquad m \neq 0,\ n \neq 0$$

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{120\pi}\left(\frac{1}{b} + \frac{2}{a}\nu_{m,0}^2\right)(1 - \nu_{m,0}^2)^{-1/2} \qquad m \neq 0,\ n = 0 \tag{10}$$

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{120\pi}\left(\frac{1}{a} + \frac{2}{b}\nu_{0,n}^2\right)(1 - \nu_{0,n}^2)^{-1/2} \qquad m = 0,\ n \neq 0$$

where $p = b/a$, $\nu_{mn} = f_{mn}/f$, and $\mathcal{R}$ is the *surface resistance* of the walls, and

$$\mathcal{R} = \sqrt{\frac{\pi\mu f}{g_m}} = \frac{2\pi\sqrt{30\mu_r}}{\sqrt{g_m\lambda_v}} = \frac{34.4\sqrt{\mu_r}}{\sqrt{g_m\lambda_v}} \tag{11}$$

In this equation, μ_r is the permeability of the walls relative to vacuum, g_m is the conductivity, and λ_v is the wavelength in vacuum for the given frequency f. For pure copper, $g = 5.80 \times 10^7$ mhos per meter and

$$\mathcal{R} = 2.61 \times 10^{-7}\sqrt{f} = 8.25 \times 10^{-7}\sqrt{\frac{f}{10}} = \frac{0.00452}{\sqrt{\lambda_v}} = \frac{0.0143}{\sqrt{10\lambda_v}} \tag{12}$$

For the TM_{mn} wave

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}(m^2b^3 + n^2a^3)}{60\pi ab(m^2b^2 + n^2a^2)}(1 - \nu_{mn}^2)^{-1/2} \tag{13}$$

The above formulas for α_m break down in the immediate vicinity of the cutoff; the attenuation does not go to infinity—it is merely large compared to the attenuation elsewhere in the pass band (see a later section on the attenuation near the cutoff).

8. CIRCULAR WAVE GUIDES

FIELDS. For the $\mathrm{TE}_{n,m}$ wave the cylindrical components of the field vectors are

$$E_\rho = A\,\frac{n}{\rho}\,J_n\left(\frac{k_{n,m}\rho}{a}\right)\sin n\varphi\, e^{-\gamma z} \qquad E_\varphi = A\,\frac{k_{n,m}}{a}\,J_n'\left(\frac{k_{n,m}\rho}{a}\right)\cos n\varphi\, e^{-\gamma z}$$

$$H_\rho = -\frac{E_\varphi}{K_z^{TE}} \qquad H_\varphi = \frac{E_\rho}{K_z^{TE}} \qquad H_z = A\,\frac{k_{n,m}^2}{j\omega\mu a^2}\,J_n\left(\frac{k_{n,m}\rho}{a}\right)\cos n\varphi\, e^{-\gamma z} \tag{14}$$

$$K_z^{TE} = \frac{j\omega\mu}{\gamma} \qquad \gamma = \sqrt{j\omega\mu(g + j\omega\varepsilon) + \frac{k_{n,m}^2}{a^2}}$$

where $k_{n,m}$ is the mth non-vanishing zero of the first derivative of the Bessel function $J_n'(k)$, and a is the inner radius of the guide. Some k's are given below:

$k_{0,1} = 3.83$	$k_{0,2} = 7.02$	$k_{0,3} = 10.17$
$k_{1,1} = 1.84$	$k_{1,2} = 5.33$	$k_{1,3} = 8.54$
$k_{2,1} = 3.05$	$k_{2,2} = 6.71$	$k_{2,3} = 9.97$
$k_{3,1} = 4.20$	$k_{3,2} = 8.02$	$k_{3,3} = 11.35$

Figure 4 shows electric lines for some TE waves. The electric lines of $\mathrm{TE}_{0,m}$ waves are circles; for this reason $\mathrm{TE}_{0,m}$ waves are called *circular electric waves.*

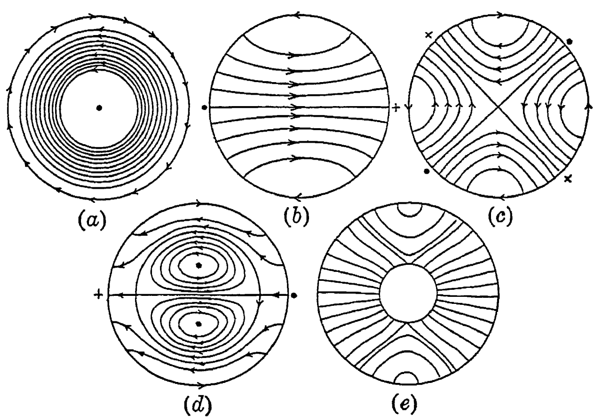


FIG. 4. Electric Lines in Circular Guides; (a) $\mathrm{TE}_{0,1}$ Wave; (b) $\mathrm{TE}_{1,1}$ Wave; (c) $\mathrm{TE}_{2,1}$ Wave; (d) $\mathrm{TE}_{1,2}$ Wave; (e) $\mathrm{TE}_{1,1}$ Wave Between Coaxial Cylinders

For the $\mathrm{TM}_{n,m}$ wave the field is

$$H_\rho = A\,\frac{n}{\rho}\,J_n\left(\frac{k_{nm}\rho}{a}\right)\sin n\varphi\, e^{-\gamma z} \qquad H_\varphi = A\,\frac{k_{nm}}{a}\,J_n'\left(\frac{k_{nm}\rho}{a}\right)\cos n\varphi\, e^{-\gamma z} \tag{15}$$

$$E_\rho = K_z^{TM} H_\varphi \qquad E_\varphi = -K_z^{TM} H_\rho \qquad E_z = -A\,\frac{k_{nm}^2}{(g + j\omega\varepsilon)a^2}\,J_n\left(\frac{k_{nm}\rho}{a}\right)\cos n\varphi\, e^{-\gamma z}$$

$$K_z^{TM} = \frac{\gamma}{g + j\omega\varepsilon} \qquad \gamma = \sqrt{j\omega\mu(g + j\omega\varepsilon) + \frac{k_{nm}^2}{a^2}}$$

where k_{nm} is the mth non-vanishing zero of the Bessel function $J_n(k)$. Some k's are given below:

$k_{0,1} = 2.40$	$k_{0,2} = 5.52$	$k_{0,3} = 8.65$
$k_{1,1} = 3.83$	$k_{1,2} = 7.02$	$k_{1,3} = 10.17$
$k_{2,1} = 5.14$	$k_{2,2} = 8.42$	$k_{2,3} = 11.62$
$k_{3,1} = 6.38$	$k_{3,2} = 9.76$	$k_{3,3} = 13.02$

Figure 5 shows magnetic lines of some TM waves. The magnetic lines of $\mathrm{TM}_{0,m}$ waves are circles; for this reason $\mathrm{TM}_{0,m}$ waves are called *circular magnetic waves.*

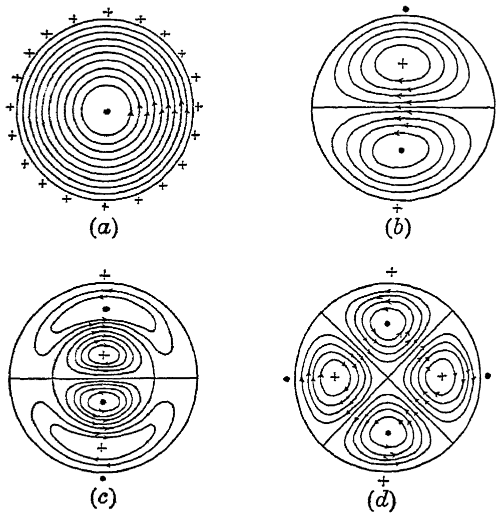


FIG. 5. Magnetic Lines in Circular Guides; (a) $TM_{0,1}$ Wave; (b) $TM_{1,1}$ Wave; (c) $TM_{1,2}$ Wave; (d) $TM_{2,1}$ Wave

CUTOFF FREQUENCIES. The cutoff frequencies and corresponding wavelengths (in vacuum) are

$$f_{nm} = \frac{ck_{nm}}{2\pi a\sqrt{\varepsilon_r}} \qquad \lambda_{nm} = \frac{2\pi a\sqrt{\varepsilon_r}}{k_{nm}} \tag{16}$$

where c is the velocity light in vacuum and k_{nm} is the quantity defined in the preceding section. For the dominant wave ($TE_{1,1}$)

$$\lambda_{1,1} = 1.7d\sqrt{\varepsilon_r} \tag{17}$$

where d is the inner diameter of the guide.

ATTENUATION IN THE PASS BAND. The propagation constant γ given in a previous section is for the case of ideal non-absorbing walls. To allow for absorption by these walls, γ should be augmented by

$$\Delta\gamma = \alpha_m(1 + j) \tag{18}$$

where α_m is the attenuation in nepers per meter caused by the absorbing walls. For TE waves

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{120\pi a}\left(\frac{n^2}{k_{nm}^2 - n^2} + \nu_{nm}^2\right)(1 - \nu_{nm}^2)^{-\frac{1}{2}} \tag{18a}$$

where $\nu_{nm} = f_{nm}/f$. The surface resistance $\mathcal{R}$ is given in the corresponding section on rectangular wave guides.

For circular electric waves $n = 0$ and

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{120\pi a}\nu_{0,m}^2(1 - \nu_{0,m}^2)^{-\frac{1}{2}} \tag{18b}$$

If the dielectric is non-dissipative, the attenuation of circular electric waves steadily decreases with increasing frequency

For dominant waves ($TE_{1,1}$)

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{a}[3.76(1 - \nu_{1,1}^2)^{-\frac{1}{2}} - 2.65(1 - \nu_{1,1}^2)^{\frac{1}{2}}]10^{-3} \quad \text{neper/meter}$$

For TM waves

$$\alpha_m = \frac{\mathcal{R}\sqrt{\varepsilon_r}}{120\pi a}(1 - \nu_{nm}^2)^{-\frac{1}{2}}$$

9. WAVE GUIDES OF ARBITRARY CROSS-SECTION

CUTOFF FREQUENCIES. For wave guides of arbitrary cross-section the cutoff wavelength (in vacuum) and the cutoff frequency are given by

$$\lambda_c = 2\pi\sqrt{\varepsilon_r}\sqrt{\frac{\iint T^2\,dS}{\iint |\text{grad } T|^2\,dS}} \qquad f_c = \frac{c}{\lambda_c} \tag{19}$$

where c is the velocity of light in vacuum, and T is proportional either to E_z or to H_z, depending on the type of the wave. This formula is exact; but generally it is not feasible to obtain T exactly. However, first-order errors in T lead to second-order errors in λ_c, and the formula is extremely useful, especially for estimating the cutoff frequency of the dominant wave; only a reasonable guess is needed for T (see the next section).

WAVE GUIDES WITH CROSS-SECTIONS OF THE TYPE SHOWN IN FIG. 6(*a*). For dominant waves in wave guides of the shape shown in Fig. 6(*a*) electric lines run across in the manner indicated in the figure. Such wave guides are essentially pairs of parallel strips, shunted on both sides with cylinders, Fig. 6(*b*). Longitudinal magnetic

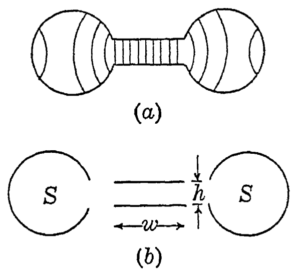


FIG. 6

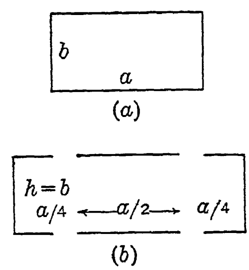


FIG. 7

lines run largely inside the cylinders, in opposite directions. On account of the increased capacitance in the middle, the cutoff frequency of such a wave guide is lower than that of a rectangular wave guide. To make an estimate of the frequency we let $T = H_z = +1$ in one cylinder and $T = -1$ in the other. Between the parallel strips we assume that the longitudinal magnetic flux varies linearly from $+1$ to -1; that is, we let $T = 1 - (2x/w)$, where w is the width of the strips.

The gradient of T is zero inside the cylinders and $(-2/w)$ between the strips. Substituting in the formula contained in the preceding section, we obtain

$$\lambda_1 \simeq \pi\sqrt{\varepsilon_r}\sqrt{\frac{2w}{h}\left(S + \frac{1}{6}wh\right)} \tag{20}$$

where h is the distance between the strips and S is the cross-sectional area of each cylinder. In the above form, the formula applies even when the cylinders are not circular. The width w should be comparable to one-half of the total width.

If this formula is applied to a rectangular wave guide whose cross-section is broken up as in Fig. 7, we find $\lambda_1 \simeq 1.81a\sqrt{\varepsilon_r}$ instead of the exact value $\lambda_1 = 2a\sqrt{\varepsilon_r}$. The error is about 10 per cent. The error diminishes as h decreases relatively to other dimensions.

10. SPECIAL CHARACTERISTICS OF WAVE GUIDES

CONDUCTION CURRENT IN THE WALLS OF A WAVE GUIDE. The conduction current in a metal wall of a wave guide is equal to the tangential magnetic intensity and is perpendicular to it. The direction is that in which a right-handed screw will advance when the H vector (which is supposed to be attached to the handle) is turned through 90° to make it coincident with the normal looking into the wall.

In the case of dominant waves in a rectangular wave guide the current in the walls parallel to the E vector is strictly transverse. In the other walls the longitudinal current is sinusoidally distributed, with the highest density in the middle; the transverse current is zero in the middle and maximum at the edges. In these walls, the transverse current flows in opposite directions from the middle. As the frequency increases, this current diminishes.

ATTENUATION IN THE VICINITY OF THE CUTOFF FREQUENCY. The following expression for the propagation constant is valid for all frequencies

$$\gamma = \sqrt{\gamma_0^2 + j\omega\mu g + 2\alpha_m\gamma_0(1 + j)} \tag{21}$$

where γ_0 is the propagation constant calculated on the assumption $g = \alpha_m = 0$. Whereas α_m is infinite at the cutoff, $\alpha_m\gamma_0$ is finite. For frequencies not too near the cutoff this formula is unnecessarily complicated, and it is recommended only for the immediate vicinity of the cutoff.

CHARACTERISTIC IMPEDANCES AND POWER TRANSFER. The ratio $K_z = E_t/H_t$ of the transverse components of the E and H vectors is called the *wave impedance*. The average power W transferred per unit area of the cross-section of the wave guide is

$$W = K_z H_{t,\text{eff}}^2 \tag{22}$$

where $H_{t,\text{eff}}$ is the effective value of H_t (the factor $1/2$ should be included if the amplitude of H_t is being used).

For the dominant wave the total power transfer P is given by

$$P = K_{P,V}V_{\text{eff}}^2 = K_{P,I}I_{\text{eff}}^2 = K_{V,I}V_{\text{eff}}I_{\text{eff}} \tag{23}$$

where V_{eff} is the effective value of the maximum transverse voltage, I_{eff} is the effective longitudinal current, and the K's are the characteristic impedances. For rectangular wave guides

$$K_{P,V} = \frac{754b}{a\sqrt{1 - \nu_{1,0}^2}} \qquad K_{P,I} = \frac{465b}{a\sqrt{1 - \nu_{1,0}^2}} \qquad K_{V,I} = \sqrt{K_{P,V}K_{P,I}} \tag{23a}$$

For circular wave guides

$$K_{P,V} = \frac{764}{\sqrt{1 - \nu_{1,1}^2}} \qquad K_{P,I} = \frac{354}{\sqrt{1 - \nu_{1,1}^2}} \qquad K_{V,I} = \sqrt{K_{P,V}K_{P,I}} \tag{23b}$$

For wave guides of the shape shown in Fig. 6, approximate formulas are

$$K_{P,V} \simeq K_{P,I} \simeq K_{V,I} \simeq \frac{377h}{w\sqrt{1 - \nu^2}} \tag{23c}$$

In these formulas ν denotes the ratio of the cutoff frequency for the dominant wave to the operating frequency.

The above expressions may be used without any reservation in power calculations. The following section should be consulted before any attempt is made to calculate reflections and standing waves.

11. WAVE-GUIDE DISCONTINUITIES

Any local change in the shape of a wave guide or any obstruction represents a *wave-guide discontinuity*. Local fields are likely to be associated with such discontinuities. A capacitor or a coil shunted across a low-frequency transmission line are examples of discontinuities. The local fields store energy during one half of the cycle and release it during the other; they act as virtual generators, operating on power borrowed from the traveling wave. If the operating frequency is higher than the cutoff frequency of the dominant wave but lower than the cutoff frequency for any other wave, there is no possibility for the energy to be sent back to its source or to the load in any mode except the dominant. In these circumstances the effect of the local field made up of higher-order waves may be represented by a reactive transducer which in its turn may be expressed as a T or pi network. The branch reactances of these equivalent networks may be computed in terms of the reflection and transmission coefficients if the latter are measured or obtained directly from a solution of an appropriate electromagnetic boundary value problem. In some special cases the equivalent network reduces to a series or shunt reactance.

If two wave guides of different dimensions are joined together, the junction introduces a reactive discontinuity of the above-mentioned type as well as a discontinuity in a characteristic impedance to the dominant wave. The reflection coefficient will be determined not merely by the impedance mismatch but by the reactive discontinuity as well. This is a general statement, and it applies to ordinary low-frequency transmission lines, for which, however, the series branches of the reactive discontinuity are so small, and the shunt branches so large, that their effect is negligible. In wave guides, on the other hand, the effect is ordinarily not negligible.

Some discontinuities are introduced inadvertently, as in joining two wave guides; others are introduced deliberately as circuit elements. Simplest circuit elements are transverse diaphragms or irises, Fig. 8(*a*), (*b*), (*c*), and transverse strips or wires, Fig. 8(*d*). If the irises and wires are thin, they are substantially shunt reactances. If the edges of the iris are perpendicular to the E vector, as in Fig. 8(*a*), the iris is capacitive; if the edges are parallel to the E vector, as in Fig. 8(*b*), the iris is inductive; in the case shown in Fig. 8(*c*), the iris is a parallel combination of an inductance and capacitance and may be designed to be an antiresonant circuit. Thin transverse wires introduce a shunt inductance nearly independent of the frequency. As the radius increases, the series branches of the equivalent transducer become important.

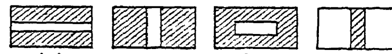

(*a*) (*b*) (*c*) (*d*)

FIG. 8

WAVE-GUIDE COMPONENTS

By George L. Ragan

12. WAVE-GUIDE CHARACTERISTICS

The form of wave guide most commonly used is a metal tube of rectangular cross-section whose width a is about twice its height b. Within a twofold frequency range, only the dominant (TE_{10}) mode is actively propagated, and the orientation of the fields within the tube is unique. That is, the electric field is perpendicular to the broader walls of the wave guide. By contrast, the frequency range within which the dominant (TE_{11}) mode in round wave guide is propagated to the exclusion of all others is only 1.31 to 1, and the orientation of the fields within the tube is not uniquely determined by the wave-guide geometry as it is in the case of rectangular wave guide. Consequently, bends and irregularities in round tubing cause changes in field orientation and even introduce elliptical polarization effects. It is because of this difficulty that round wave guide is little used as a transmission line. However, round wave guide is frequently used in short sections for rotary joints. In this application, the symmetrical TM_{01} mode is used.

Expressions for calculating cutoff wavelength, attenuation in copper wave guide, and power transmitted in common wave-guide modes appear in Table 1.

Table 1. Wave-guide Expressions *

Wave-guide Shape	Mode	λ_c	α/α_1†	P/P_1‡
Rectangular..........	TE_{10}	$2a$	$4.16\left[\frac{a}{2b} + \left(\frac{\lambda}{\lambda_c}\right)^2\right]$	$6.63ab$
Round..............	TE_{11}	$1.706d$	$3.55\left[0.420 + \left(\frac{\lambda}{\lambda_c}\right)^2\right]$	$4.98d^2$
Round..............	TM_{01}	$1.306d$	2.72	$2.83d^2\left(\frac{\lambda_c}{\lambda}\right)^2$ if $\lambda > 0.657d$

* Tables and figures in articles 12–19 are reproduced by permission from G. L. Ragan, *Microwave Transmission Circuits*, Vol. 9, Rad. Lab. Series, McGraw-Hill Book Co., 1948.

† α = attenuation for copper wave guide, decibels per meter.

$$\alpha_1 = \frac{10^{-4}(\lambda/\lambda_c)}{\lambda^{3/2}\sqrt{1-(\lambda/\lambda_c)^2}}$$

‡ $P_1 = 10^{-4}\,E_{\max}{}^2\sqrt{1-\left(\frac{\lambda}{\lambda_c}\right)^2}.$

The notation used in this article is as follows:

a = larger inside dimension of rectangular wave guide, in meters.
b = smaller inside dimension of rectangular wave guide, in meters.
d = inside diameter of round wave guide, in meters.
λ = free-space wavelength, in meters.
λ_c = cutoff wavelength for the mode, in meters.
α = attenuation in decibels per meter.
P = power transmitted, in watts.
E = electric field intensity, in volts per meter.

In Table 2, numerical values for a few representative wave guides are given. The figure E_{max} = 30,000 volts/cm is based on experimental work on air at atmospheric pressure (M.I.T. *Radiation Laboratory Series*, McGraw-Hill Book Co., Vol. 9, Chapter 4).

Table 2. Characteristics of Representative Wave Guides

(Courtesy McGraw-Hill Book Co.)

Army-Navy Designation	Guide Size OD, in.	Wall, in.	Wave-length, cm	Power,* mega-watts	Loss,† db/m Copper	Wavelength ‡ Band, cm
A. Rectangular (TE_{10} Mode)						
RG-48/U.....	3 × 1.5	0.080	10.0	10.5	0.0199	7.3 – 13.0
RG-51/U.....	1.25 × 0.625	0.064	3.2	1.77	0.0725	2.9 – 5.1
RG-52/U.....	1.0 × 0.5	0.050	3.2	0.99	0.117	2.3 – 4.1
None........	0.5 × 0.25	0.040	1.25	0.223	0.346	1.07 – 1.9
B. Round (TE_{11} Mode)						
None........	3 ID	.....	10.0	16.6	0.0140	10.0 – 11.7
None........	1 OD	0.032	3.2	1.57	0.0847	3.18 – 3.64

* Calculated assuming E_{max} = 30,000 volts/cm.
† Calculated values for copper.
‡ Based on maximum wavelength 10 per cent below cutoff wavelength, minimum wavelength 1 per cent above cutoff wavelength of next higher mode.

13. FLEXIBLE WAVE GUIDES AND COUPLING UNITS

Flexible wave guide is used similarly to flexible coaxial cable. Applications include connections to shock-mounted units; connections to pieces of equipment which must be moved about; temporary or emergency replacement lines; and as patch connections, particularly in test equipment.

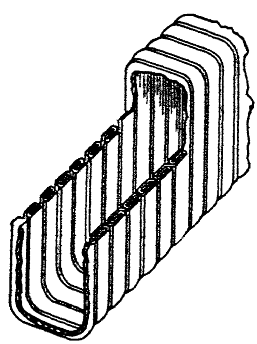


Fig. 1. Wound Metal-hose Wave Guide (Courtesy McGraw-Hill Book Co.)

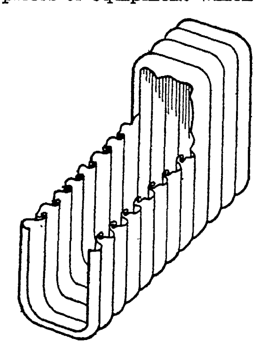


Fig. 2. Titeflex Wave Guide (Courtesy McGraw-Hill Book Co.)

Two types of construction which have proved to be especially useful are illustrated in Figs. 1 and 2. Figure 1 shows the "metal-hose" type manufactured by the American Metal Hose branch of the American Brass Co., Waterbury, Conn. This type is made of fairly heavy metal wound in the interlocking manner indicated. Flexibility is afforded by sliding of the interlocked contacting surfaces. The "Titeflex" type shown in Fig. 2 is manufactured by Titeflex, Inc., Newark, N. J. Titeflex is made of thin metal wound as shown and soft-soldered. Flexibility is afforded by flexure of the thin metal convolutions. It has been found that a molded-on rubber jacket affords needed protection to both types and in addition improves the performance of the metal-hose type by causing lower contact resistance. A complete rubber-covered section is shown in Fig. 3. Such wave guides may be bent on radii equal to about 20 times the respective nominal dimensions of the wave guide.

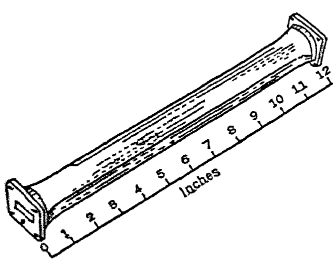


Fig. 3. Rubber-covered Flexible Waveguide Assembly (Courtesy McGraw-Hill Book Co.)

Several other types of flexible wave guide have been found useful, especially in short lengths, as flexible coupling units. These include: (*a*) Corrugated wave guide; essentially a thin-walled metal bellows of rectangular cross-section (American Metal Hose). The depth of convolutions must be a small fraction of a wavelength. (*b*) Spun bellows assembly; a soft-soldered assembly of thin-walled units each of which includes a circular bellows section the depth of whose convolution is effectively one-half wavelength (American Metal Hose).

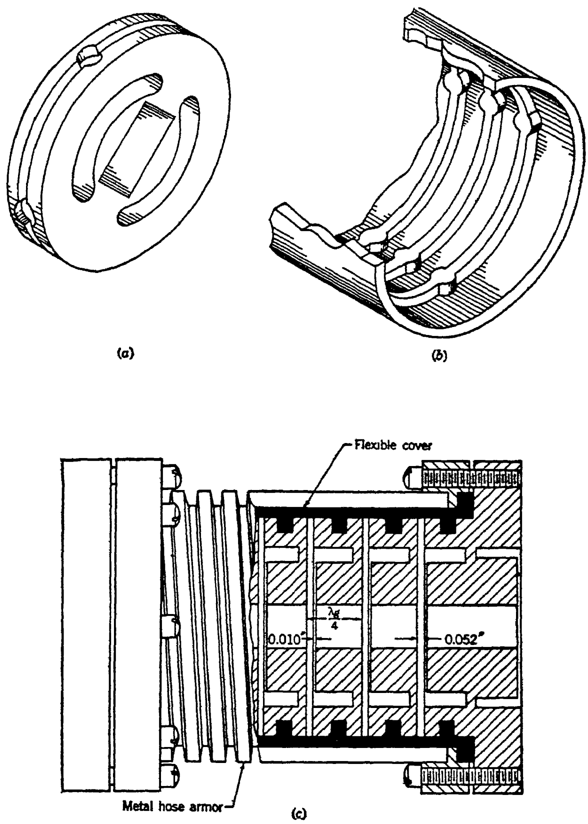


FIG. 4. Typical Vertebral Assembly; (*a*) is a Single Choke Disk; (*b*) is the Flexible Cover; (*c*) is the Assembly (Courtesy McGraw-Hill Book Co.)

(*c*) Cook bellows assembly (Cook Electric Co., Chicago, Ill.); a somewhat more rugged and broader-band bellows similar in principle to (*b*).

A coupling unit providing a maximum flexibility in longitudinal, transverse, and angular displacements is the so-called vertebral type illustrated in Fig. 4. This unit is based on the choke-flange or capacity-type wave-guide connector described in article 14 below. Power leakage from the open space between adjacent sections is minimized by the action of the choke grooves, and fairly large displacements may be tolerated without causing serious impedance mismatch.

14. WAVE-GUIDE CONNECTORS

Two sections of wave guide are usually joined by couplings of either contact type, Fig. 5, or choke-flange type, Fig. 6. The contact type has some advantages and provides excellent results if certain precautions in design and use are observed. The choke-flange type, however, is found to be more reliable for general use.

The design of contact couplings must be such that good contact is secured at all points around the periphery of the butt-joined wave guides. Particular care must be exercised to avoid the formation of cavities by permitting the contact to be made at points on the bolting flanges before the wave-guide ends are forced into good contact. The contact surfaces must be carefully machined and must be kept free of dirt, corrosion, and mechanical deformations.

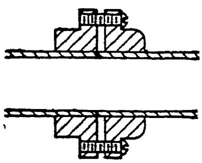


FIG. 5. Contact Coupling (Courtesy McGraw-Hill Book Co.)

When these precautions are observed, a junction is obtained which is practically perfect in all respects. This type of coupling is exceedingly valuable in certain laboratory design work where the required care in use can be taken. Likelihood of deterioration of the quality of the contacts in use under field conditions, however, presents a serious obstacle to their general utility. One contact-type coupling which has given satisfactory service for use with 1 1/2 in. by 3 in. wave guide is that designated by the Army-Navy Cable Coordinating Committee as UG-65/U and UG-66/U.

The choke-flange type coupling of Fig. 6, in which quality of contact is unimportant, affords a connector admirably suited to use under adverse conditions. In addition, such a coupling scheme is very useful in rotary joints. Parts (*b*) and (*c*) of Fig. 6 are designed principally for such use. The wave-guide ends are slightly separated, and currents interrupted by the gap *A* excite a folded section of line which surrounds the gap. This section of line is terminated in a short circuit (closed end), and its length is effectively one-half wavelength. Hence it presents at its input end essentially zero impedance to the flow of currents interrupted by the gap *A*. The contact between flanges occurs at the point *B*, which is at the midpoint of the half-wavelength line, and hence at a point of essentially zero current. It is for this reason that the quality of the contact between flanges is unimportant.

Naturally, there is only one specific wavelength for which the conditions outlined above exist, and at this wavelength only is the connector perfectly matched (i.e., reflectionless). In the design of such connectors, one must be guided by the principles discussed below if a low degree of frequency sensitivity is to be achieved. Since the same principles are involved in the choke-type coupling scheme commonly used in most rotary joints (see article 17 below), they will be discussed in some detail.

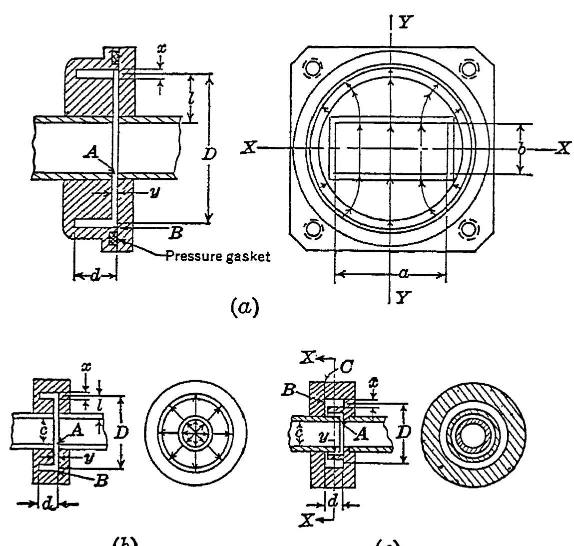


FIG. 6. Choke-flange Couplings; (*a*) Rectangular Wave Guide; (*b*) and (*c*) Circular Wave Guide (Courtesy McGraw-Hill Book Co.)

In considering the action of the choke-flange coupling, it is convenient to consider separately two sections, each effectively one-quarter wavelength long at the design frequency: (*a*) the radial section from *A* to *B*, and (*b*) the circular groove of depth *d*. The circular groove is terminated in zero impedance and presents at *B*, an effective quarter wavelength away, a high impedance which, in series with the contact resistance at *B*, terminates the radial section of line. This high-impedance termination of the radial line is transformed to a low impedance at *A*, an effective quarter wavelength away. The design problem resolves itself into that of providing, at frequencies in the neighborhood of the design frequency, a maximum impedance at *B* and a minimum impedance at *A*. It is easily argued that these objectives are realized (*a*) by maximizing the characteristic impedance of the groove section by making the groove width x as large as is practical, and (*b*) by minimizing the characteristic impedance of the radial section by making the gap y as small as is practical.

For example, let us compare the voltage standing wave ratio (*VSWR*) introduced to $\lambda = 9$ cm by two chokes, both perfectly matched at 10.7 cm, differing only in regard at the values of x and y. For one design, $x = 0.150$ in., $y = 0.104$ in., $VSWR = 1.13$. For the other design, $x = 0.250$ in., $y = 0.050$ in., and $VSWR = 1.03$.

Design details and performance figures for a number of choke-flange units are given in Table 3.

Table 3. Choke-coupling Design Details
(Courtesy McGraw-Hill Book Co.)

Army-Navy Type Choke Flange	Guide Dimensions, in.		Choke Dimensions, in.				Design Wave-length, cm	Band Width for $r = 1.05$, %
			D	x	y	d		
	a	b	Rectangular-wave-guide Choke. Fig. 6 (*a*)—TE_{10} Mode					
UG-54/U-53/U	2.84	1.34	4.015	0.250	0.050	1.120	10.7	±15
U-200/U-214/U	2.84	1.34	3.75	0.250	0.030	0.865	9.0	±15
UG-40/U-39/U	0.90	0.40	1.183	0.063	0.031	0.347	3.20	±6
UG-52/U-51/U	1.125	0.50	1.332	0.063	0.031	0.347	3.20	....
None	0.900	0.40	1.155	0.125	0.010 *	0.355 *	3.30	>±6
UG-117/U-116/U	0.420	0.170	0.501	0.029	0.008	0.137	1.25	>±2
None	0.420	0.170	0.589	0.063	0.008	0.156	1.25	>±4
	c		Circular-wave-guide Chokes. Fig. 6 (*b*)—TM_{01} Mode					
None	0.4675		0.713	0.050	0.015	0.153	1.25	>±4
None	1.187		1.479	0.093	0.030	0.312	3.30	>±6

* Designed for 0.115-in. separation between choke and flange.

15. BENDS, TWISTS, AND ANGLES

If rectangular wave guide is bent or twisted gradually enough, i.e., in a length representing several wavelengths, the reflections set up in the wave guide are negligibly small. Alternatively, it may be bent or twisted rather abruptly without introducing serious reflections if the mean length of bend or twist is a multiple of half the guide wavelength. Care must always be exercised to avoid the introduction of bumps or ripples in the wave-guide walls or of excessive distortions of the cross-sectional dimensions. The achievement of this end is materially assisted by filling the wave guide with a low-melting-point alloy such as Woods metal or Cerrobend before working it, and melting it out afterwards.

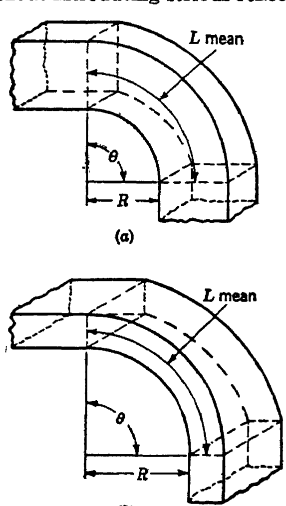


Fig. 7. Wave-guide Bends; (*a*) H Bend; (*b*) E bend (Courtesy McGraw-Hill Book Co.)

Figure 7 illustrates two types of bend, designated as *H*-plane and *E*-plane bends, depending on whether the bending radius vector lies in the plane of the magnetic field lines (*H*) or electric field lines (*E*). Incidentally, these bends are sometimes conveniently referred to as "hard bends" (*H*) or "easy bends" (*E*) for obvious mechanical reasons. The performance data for some bends and twists are given in Table 4.

It is sometimes preferable, in order to achieve minimum space factor, to substitute fabricated corners for bends. Two types of well-matched corner are illustrated in Fig. 8. The double-mitered type is usually to be preferred (*a*) because the dimensional tolerances are larger, and (*b*) because this type can carry more power, since it has less acute corners and a less restricted cross-section. Both types may be designed for any desired angle, not being restricted to the 90° type drawn.

It is found experimentally that the mean separation L between miters in the double-mitered *E*-plane corner is a quarter guide wavelength for best match. This is easily understood as providing a cancellation of mismatches by reason of quarter-wave spacing between identical reflections. In double-mitered *H*-plane corners, however, the spacing deviates slightly from the expected quarter-wave value. This is presumably because of the disturbance due to a greater

Table 4. Performance of Wave-guide Circular Bends and Twists
(Courtesy McGraw-Hill Book Co.)

Type	Wave-guide Size ID, in.	Inside Radius R, in.	θ	Design Wave-length, cm	Band Width for r below 1.05, %
E-plane bend	1.34 × 2.84	6	45°	10	$> \pm 20$
	1.34 × 2.84	6	90°	10	$> \pm 20$
	1.125 × 0.50	2	90°	3.3	$\pm$ 9
	0.90 × 0.40	0.50	180°	3.3	$> \pm$ 9
	0.90 × 0.40	0.25	90°	3.3	$> \pm$ 9
	0.90 × 0.40	3.00	90°	3.3	$> \pm$ 9
	0.420 × 0.170	0.50	90°	1.25	$> \pm$ 4
H-plane bend	1.34 × 2.84	6	45°	10	± 10
	1.34 × 2.84	6	90°	10	$> \pm 20$
	1.125 × 0.50	2	90°	3.3	$> \pm$ 9
	0.90 × 0.40	0.192	90°	3.35	$\pm$ 4
	0.90 × 0.40	1.1875	90°	3.30	$\pm$ 9
	0.420 ± 0.170	0.50	90°	1.25	$> \pm$ 4
Type	**Wave-guide Size ID, in.**	**Length, in.**		**Design Wave-length, cm**	**Band Width for r below 1.05, %**
Twists	0.900 × 0.400	2		3.4	± 6
	0.900 × 0.400	3		3.4	± 3.7
	1.125 × 0.500	4		3.3	$> \pm 9$
	0.420 × 0.170	1 1/4		1.25	$> \pm 4$
	0.420 × 0.170	2 1/2		1.25	$> \pm 4$

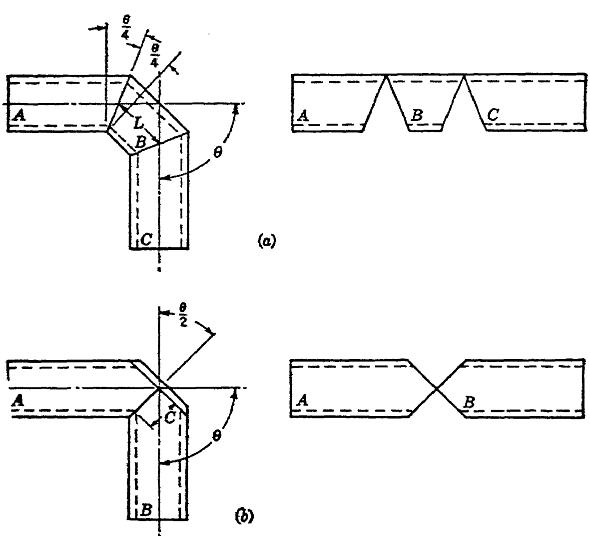


FIG. 8. Wave-guide Corners; (a) Double-mitered Type; (b) Single-miter, Cutoff Type (Courtesy McGraw-Hill Book Co.)

excitation of local fields. Figure 9 gives the spacing found experimentally to give well-matched double-mitered H-plane corners of 90° total angle. Her λ_0/λ_c is the ratio of design free-space wavelength to the cutoff wavelength. The band width (for $VSWR$ below 1.06) is found to range from about 20 per cent (occurring for λ_0/λ_c of about 0.6) to about 8 per cent (for λ_0/λ_c of about 0.85).

16. IMPEDANCE MATCHING AND IMPEDANCE TRANSFORMERS

In designing a specific wave-guide component, certain parameters are varied in an attempt to arrive at a design which performs its specified function and at the same time introduces into the wave guide the minimum impedance discontinuity, i.e., the minimum of reflection of the incident r-f wave. It is frequently impossible to achieve the desired degree of freedom from reflected waves without sacrificing quality of performance in some other respect. When such an occasion arises, it is necessary to compensate for this im-

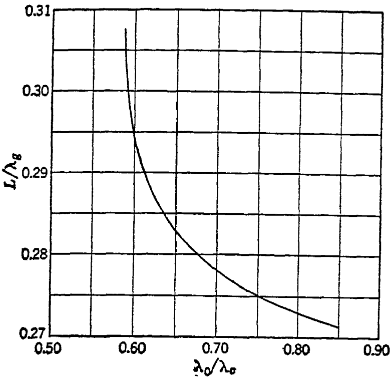


FIG. 9 Design Curve for Double-mitered H-plane Corners (Courtesy McGraw-Hill Book Co.)

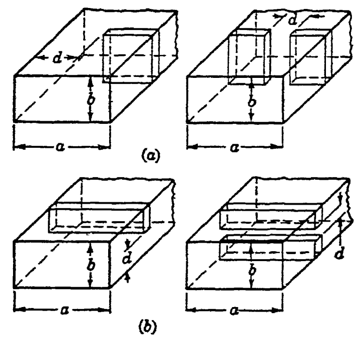


FIG. 10. Wave-guide Impedance-matching Diaphragms; (a) Inductive Types; (b) Capacitive Types (Courtesy McGraw-Hill Book Co.)

pedance discontinuity by introducing into the wave guide an impedance-matching transformer.

Impedance transformers may be classified into two general categories, fixed and variable. The fixed type is inserted at the time of fabrication according to instructions given by the designer, whereas the variable type may be altered by the user of the equipment to achieve the desired performance. Though the variable type may give superior performance when adjusted with the required skill and care, it is capable of doing more harm than good if improperly adjusted. The variable type is, therefore, to be avoided whenever possible.

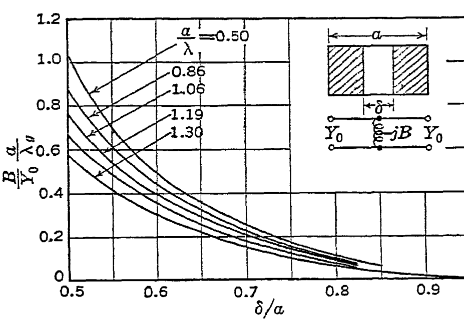


FIG. 11. Design Curves for Symmetrical Inductive Diaphragms (Courtesy McGraw-Hill Book Co.)

The fixed-type impedance transformer usually consists of one or more thin metal strips soldered into the wave guide in one of the forms shown in Fig. 10. The most widely used is the symmetrical inductive diaphragm shown in the upper right figure. The equivalent circuit and design data for this type are given in Fig. 11. The asymmetrical inductive diaphragm, design data for which are given in Fig. 12, is used somewhat less, as the local field distortions set up by it are more extensive. The capacitive types are little used because of the restriction of the permissible power level imposed by the high electrical fields occurring across the gap formed.

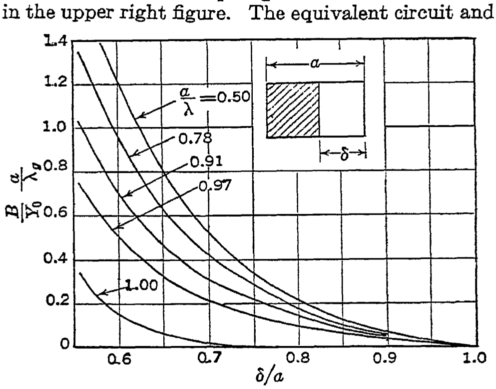


FIG. 12. Design Curves for Asymmetrical Inductive Diaphragms (Courtesy McGraw-Hill Book Co.)

The curves of Figs. 11 and 12 are for infinitely thin metal strips. A rough compensation for increased thickness may be made by assuming a strip to extend into the open space by an amount equal to half the thickness of the strip.

The normalized susceptance B/Y_0 required to correct for a measured voltage-standing-wave ratio r may be determined by the relation

$$\frac{B}{Y_0} = \frac{r-1}{\sqrt{r}} \tag{1}$$

The distance d, from the determined position of a minimum in the voltage-standing-wave pattern, toward the load (for inductive diaphragms) or toward the source (for capacitive diaphragms) may be determined from the relation

$$\frac{d}{\lambda_g} = \frac{90° - \tan^{-1}(1/2\, B/Y_0)}{720°} \tag{2}$$

Alternatively, both B/Y_0 and d may be read from a transmission line chart or admittance diagram.

Several other types of fixed-impedance transformer have seen limited use. One of these is the quarter-wave transformer of ordinary transmission lines. A wave-guide section of the desired lower characteristic impedance is formed by soldering a plate of suitable thickness to one of the two broad walls of the guide. This plate is the full wave-guide width and a quarter guide wavelength long. Since guide width a is not changed, characteristic impedance is proportional to unfilled wave-guide height b.

Another scheme utilizes capacitive "buttons" soldered in place, or dents formed by a suitable rounded tool. Either button or dent forms a projection from the broad wall of the guide which acts essentially as a shunting capacitive element.

Variable-impedance transformers or "tuners" have appeared in numerous forms. They have usually either two or three adjustable elements appropriately spaced along the guide. Among the forms which the adjustable elements have assumed are: (*a*) "Stubs," or branching sections of wave guide perpendicular to the main wave guide and containing adjustable short-circuiting plungers. The stub line may branch from either the broad wall (E-plane stub) or narrow wall (H-plane stub) of the guide. (*b*) Screws, inserted through threaded holes in the broad wall. (*c*) "Slugs," obstacles of either dielectric or metallic materials, designed to alter the characteristic impedance of the guide in the section into which they are inserted (usually a quarter guide wavelength long). These slugs are ordinarily inserted through a narrow longitudinal slot in one of the broad wave-guide walls.

Stubs may introduce either inductive or capacitive susceptance. Three stubs spaced at quarter-wavelength intervals can, in theory, transform any load impedance into any desired input impedance. In practice, if one is interested in matching out or introducing a *VSWR* of 2 or less, two stubs spaced an odd number of eighth wavelengths apart are adequate and have the advantage of easier adjustment.

Small-diameter screws inserted into the wave guide introduce a capacitive susceptance only. This limitation makes the attainment of a proper adjustment much more difficult than that for stubs. As the screw diameter is increased to a size comparable to the wave-guide dimensions, it achieves an inductive effect when retracted, just as does a stub, which it begins to resemble. This opens up the possibility of a screw-type tuner which is relatively easily adjusted. As for stub tuners, an odd number of eighth wavelengths proves to be a good spacing.

A single screw of small diameter, mounted on a sliding sleeve fitted closely around the wave guide, and projecting through a longitudinal slot in the wave guide, constitutes a very useful tuning device. By adjusting both insertion length and position along the slot, any tuning requirement is easily met.

"Slug" tuners may be similar in action to the single screw tuner just described. Or they may consist of two identical slugs whose overall reflection is varied by varying the spacing between them, and the phase of this overall reflection is varied by sliding the two slugs, as a single unit, along the guide.

A simple form of variable-impedance transformer is the phase shifter or "line stretcher" type. This type does not alter the magnitude of the wave reflected by an impedance discontinuity but merely alters its phase. Such a device is very useful in promoting stability of magnetron oscillators in long lines.

One common form is made simply by cutting longitudinal slots in both broad walls of the wave guide and squeezing the section thus formed to alter the wave-guide width. As width changes, guide wavelength and hence total phase length of the section change.

17. TRANSITION UNITS

Whenever it is desired to couple two different wave guides together, either in the same or different modes, some type of transition unit is needed. Similarly, a transition between coaxial lines and wave guides is frequently needed. One very important need for transition units is in connection with rotary joints, where the dominant wave-guide mode cannot be used (see article 18).

A number of types of transitions between coaxial lines and wave guides are illustrated in Fig. 13. Types (*e*) and (*f*) are especially suited for low-power work; types (*g*), (*d*), and (*c*) are recommended for intermediate powers; and types (*h*) and (*i*) are especially suitable for high powers.

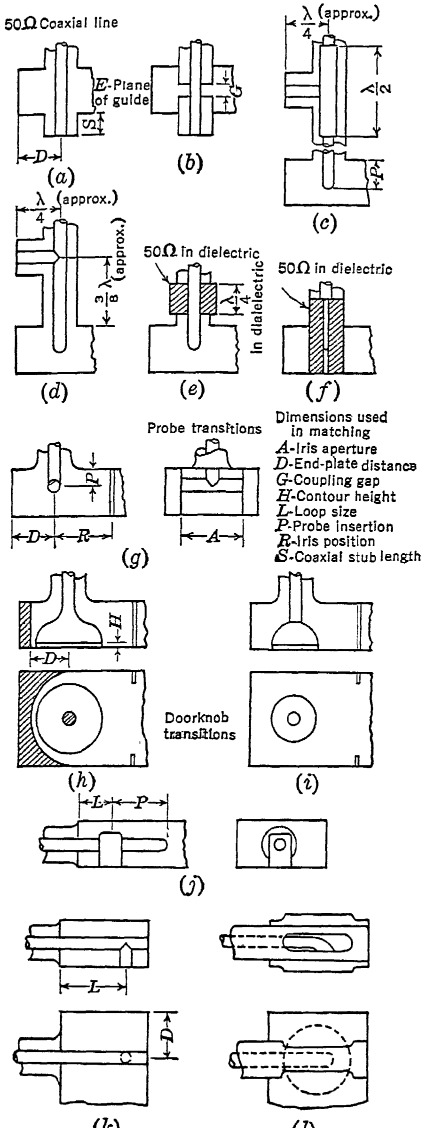


FIG. 13. Transitions from Coaxial Line to Wave Guide (Courtesy McGraw-Hill Book Co.)

A simple transformer designed to couple from 1 by 1/2 in. wave guide to 1 1/4 by 5/8 in. wave guide is shown in Fig. 14. Such a transformer section may be calculated by making

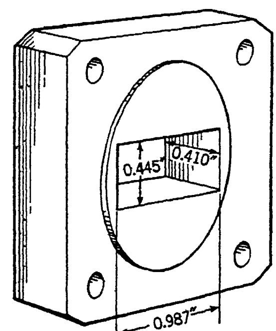


FIG. 14. Quarter-wavelength Transformer between 1 x 1/2 inch to 1 1/4 x 5/8 inch Wave Guides (Courtesy McGraw-Hill Book Co.)

the intermediate section a quarter wavelength long (in terms of its own λ_g), and choosing its dimensions so as to give a characteristic impedance equal to the geometric mean of those of the joined wave guides. For this calculation, the characteristic impedance may be taken as

$$Z_0 = \sqrt{\frac{\mu}{\varepsilon}}\,\frac{\lambda_g}{\lambda_0}\,\frac{b}{a} \tag{3}$$

If one prefers, two rectangular wave guides could be joined by a relatively long tapered

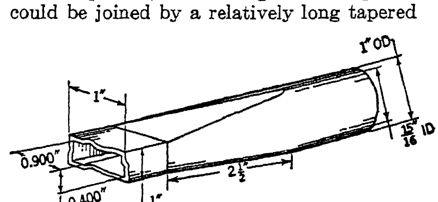


FIG. 15. Taper from Rectangular to Round Wave Guide (Courtesy McGraw-Hill Book Co.)

section. Such a taper is commonly used between rectangular and round wave guides, as illustrated in Fig. 15. If space does not permit such a taper between rectangular and round guides, a transformer section such as that shown in Fig. 16 may be used.

Transitions between the TE_{10} mode in rectangular wave guide and the TM_{01} mode in round wave guide are particularly useful in rotary joints. Three such transition units are shown in Figs. 17, 18, and 19. Varying degrees of complexity are illustrated. In Fig. 17, the parameters were adjusted so that satisfactory impedance match and TM_{01} mode purity are obtained without the matching diaphragms or mode-filter ring shown in the other designs. The design of Fig. 18 is especially recommended for high-power work, as no sharp corners are present to cause breakdown.

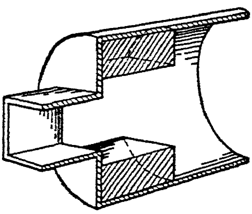


FIG. 16. Quarter Wavelength Transformer between Round and Rectangular Wave Guides (Courtesy McGraw-Hill Book Co.)

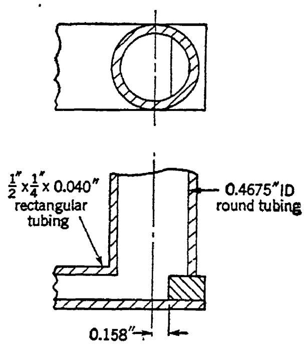


FIG. 17. Transition from Rectangular TE_{10} to Round TM_{01} Mode for 1.25 cm Wavelength (Courtesy McGraw-Hill Book Co.)

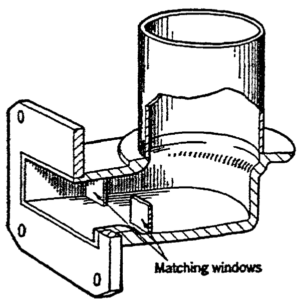


FIG. 18. TM_{01} Transition for High-power Use (Courtesy McGraw-Hill Book Co.)

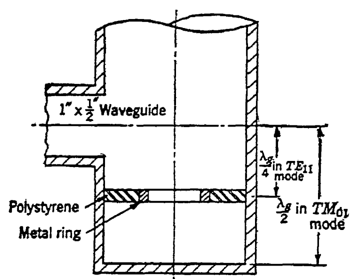


FIG. 19. TM_{01} Transition with Combination Stub and Resonant Ring Filter (Courtesy McGraw-Hill Book Co.)

18. MOTIONAL JOINTS

This classification includes rotary joints, oscillating joints, hinge or "knuckle" joints, and universal joints. Rotary joints, exemplified by Figs. 20, 21, and 22, permit continued rotation about an axis; oscillating joints permit limited rotary oscillations about an axis.

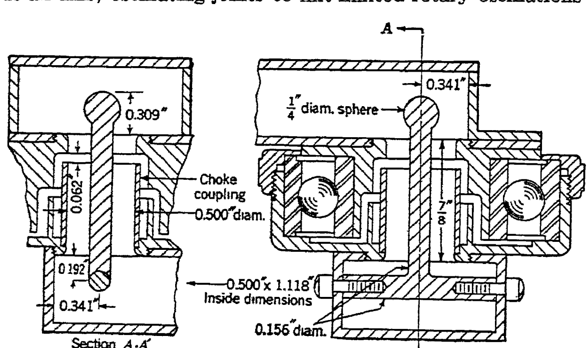


FIG. 20. Rotary Joint, Combining Coaxial Line and Wave Guide, for Use at 3-cm Wavelength (Courtesy McGraw-Hill Book Co.)

Hinge or knuckle joints, Fig. 23, permit angular displacements of one wave guide with respect to another, the axis of the hinging being perpendicular to the wave-guide axis. Universal joints, Fig. 24, permit the motion provided by gimbals.

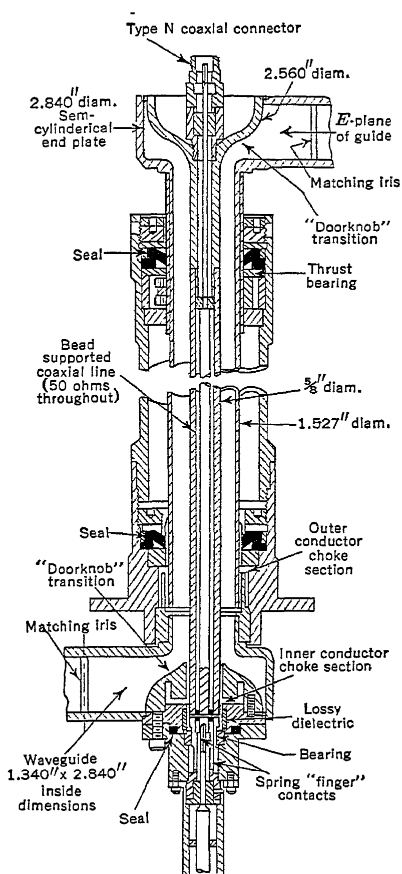


FIG. 21. Rotary Joint for 10-cm Wavelength (Courtesy McGraw-Hill Book Co.)

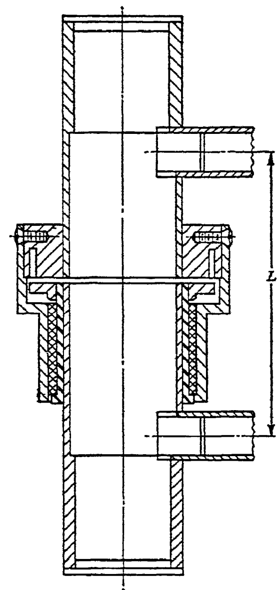


FIG. 22. Rotary Joint Using TM_{01} Mode in Round Wave Guide (Courtesy McGraw-Hill Book Co.)

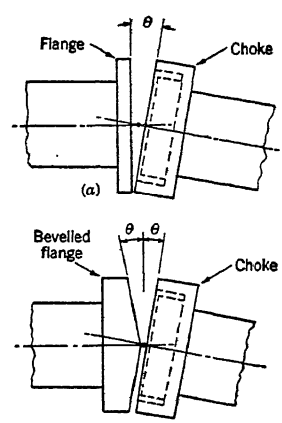
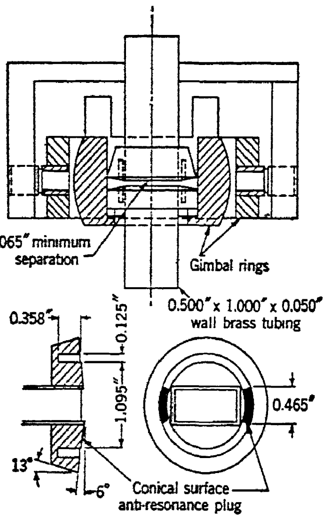


FIG. 23. Hinge or Knuckle Joints (Courtesy McGraw-Hill Book Co.)

Any transition unit from the dominant wave-guide mode to a coaxial line or to a round wave guide excited in a circularly symmetrical mode may serve as the basis for a rotary joint. In Figs. 20 and 21, the actual rotation is accomplished in coaxial line. Gaps in the coaxial line conductors are bridged by "choke couplings" similar to those described in article 14. A convenient feature of the high-power rotary joint of Fig. 21 is the inclusion, within the center conductor of the main joint, of a second coaxial rotary joint for an auxiliary transmission line.

An important consideration in the design of rotary joints using the TM_{01} mode is concerned with the choice of a length L, Fig. 22, which avoids troublesome resonance effects. Such resonances are associated with the dominant mode fields which inevitably exist, despite efforts to avoid their excitation, in the round wave-guide section.

Any rotary joint may, obviously, be used as an oscillating joint merely by restricting its rotational amplitude. A simpler design, however, may be arrived at by applying the principles of the vertebral assembly of chokes and flanges described in article 13.



Two constructions of H-plane hinge joints are illustrated in Fig. 23. Again, the choke and flange connector is used as a basis of the design. Similar hinge joints of E-plane type are in use.

The universal type joint of Fig. 24 is essentially a double hinging obtained by means of gimbals and suitable modifications of the choke-flange principle. In this design, two chokes are opposed, rather than choke and flange, in order to permit greater freedom of motion without leakage of power from the openings around the joint. The antiresonance plugs, shown darkened, prevent trouble from resonances which are found to occur whenever two chokes are opposed. The same resonance trouble is found when wave guides are joined by opposing two chokes.

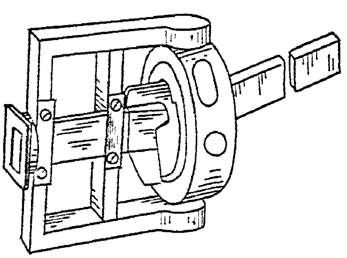


FIG. 24. Universal Joint (Courtesy McGraw-Hill Book Co.)

19. OTHER COMPONENTS

The foregoing discussion is, of course, far from exhaustive. Space has permitted the inclusion of only a few designs representative of the types discussed. In addition, several types of component have not been treated, even cursorily, because of space limitations.

Among the more important omissions are the so-called T-R or duplexing components which permit the transmission of r-f power and the reception of r-f signals on the same wave-guide and antenna assembly. Still other circuits, which are conspicuously omitted, are the so-called mixer circuits in which received signals are mixed in suitable crystals or tubes with a local oscillator signal to generate the i-f signals supplied to the receiver.

Those interested in pursuing these omitted items will find them described in the appropriate volume of the M.I.T. *Radiation Laboratory Series* listed in the bibliography.

BIBLIOGRAPHY

1. Books of the Massachusetts Institute of Technology *Radiation Laboratory Series*, published by McGraw-Hill Book Co., New York, 1947 and 1948.
   (a) Vol. 9, *Microwave Transmission Circuits;* Ragan, editor.
   (b) Vol. 17, *Components Handbook;* Blackburn, editor.
   (c) Vol. 8, *Principles of Microwave Circuits;* Montgomery, Dicke, and Purcell, editors.
   (d) Vol. 10, *Waveguide Handbook;* Marcuvitz, editor.
   (e) Vol. 14, *Microwave Duplexers;* Smullin, editor.
   (f) Vol. 16, *Microwave Mixers;* Pound, editor.
   (g) Vol. 1, *Radar Systems Engineering;* Ridenour, editor.
2. *Microwave Transmission Design Data*, Sperry Gyroscope Co., May 1944.
3. *Reference Data for Radio Engineers*, Federal Telephone and Radio Corp., Second Edition, 1946.
4. *Microwave Techniques*, prepared by M.I.T. Radiation Laboratory, Bureau of Ships Publication 900,028 (1944).

TRANSMISSION IN SPACE

By J. C. Schelleng

20. WAVE PROPAGATION AND GENERAL VIEW OF THE RADIO SPECTRUM

Radio propagation is a special instance of the propagation of electromagnetic waves. As electromagnetic waves have so much in common with light, an understanding of radio propagation begins with an understanding of optics. Radio propagation provides examples of most optical phenomena: interference, reflection, simple refraction and double refraction, diffraction, etc. Many of the standard formulas of optics can be carried over without change into the radio field. Although it is beyond the scope of this article to discuss these fundamental considerations, a few general principles will be mentioned.

RADIATION AND TRANSMISSION IN FREE SPACE. Radiation of electromagnetic waves take place whenever an electrical charge is accelerated. The wave which is set up is transverse, its electrical field at any point being a vector perpendicular to the direction of transmission, lying in the plane specified by the direction of propagation and the direction of acceleration of charge, and measured in units of potential per unit of length in the direction of the field, e.g., microvolts per meter. When the space surrounding the source is free of material, the electrical field is propagated outward with the velocity of light, and at sufficiently great distances it is proportional to that component of the acceleration which is parallel to the electric field at the point in space under consideration (Fig. 1). These statements are true whether the acceleration is sinusoidal or not. In radio communication, the accelerated charges are the electrons in the conductors of the antenna. For engineering purposes it is more convenient to use formulas involving quantities other than acceleration and charge. Thus, in free space, the electric field intensity, E, from a short electric dipole is

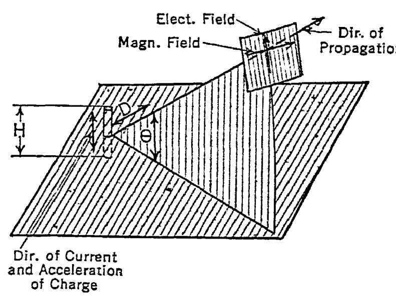


FIG. 1. Vector Relations in Radiation

$$\mathcal{E} = 60\pi \frac{hI}{\lambda d} \cos\theta \tag{1}$$

h, d, and λ are the effective height, distance from doublet to measuring point and wavelength, all in the same unit of length (e.g., meters); $\mathcal{E}$ is in volts per unit length (e.g., per meter), and I is in amperes; see article 28, pp. 5–52. A convenient form for either electric or magnetic dipole is

$$\mathcal{E} = \frac{\sqrt{45P}}{d} \cos\theta \tag{2}$$

in which P is the radiated power in watts, and $\mathcal{E}$ and d are in the same units as for eq. (1). The field varies inversely with the distance. For an especially useful formula for the ratio of power picked up by a receiving antenna to that radiated from a distant transmitter, the conditions being those of free-space transmission, the reader is referred to Section 6, article 28, eq. (11).

GROUND WAVE AND SKY WAVE. Two general modes of wave propagation are useful in radio communication: the *ground wave*, which passes *along* the surface of the earth; and the *sky wave*, which, traveling at an angle with the surface, passes through the lower atmosphere, is reflected from the upper atmosphere, and is enabled in this way to return to the earth at a distant point. The ground wave is used over short distances; the sky wave, or ionospheric wave, is required for the longer spans. Intermediate ranges may involve either or both, depending on the frequency used, the time of day, and other circumstances.

The reciprocity theorem of Lord Rayleigh, originally derived and widely used for electric-circuit analysis, has been shown by Carson and others to be true in cases involving radiation (*B.S.T.J.*, April 1930, and earlier papers). It results from one form of this theorem that with certain limitations the efficiency of radio transmission in opposite directions is the same, provided that the usual measures have been taken to match the generator and load to the antennas, and that the path is free of elements which fail to act reciprocally by themselves. A one-way amplifier is an obvious example of such an element. Another example, less obvious, actually occurs in the upper atmosphere itself, namely, the ions which tend to spiral about the earth's magnetic field in one direction but not in the other, thus producing a non-reciprocal element. As a result, strict reciprocity can be expected where the ground wave is concerned, but it becomes doubtful with the sky wave, and it

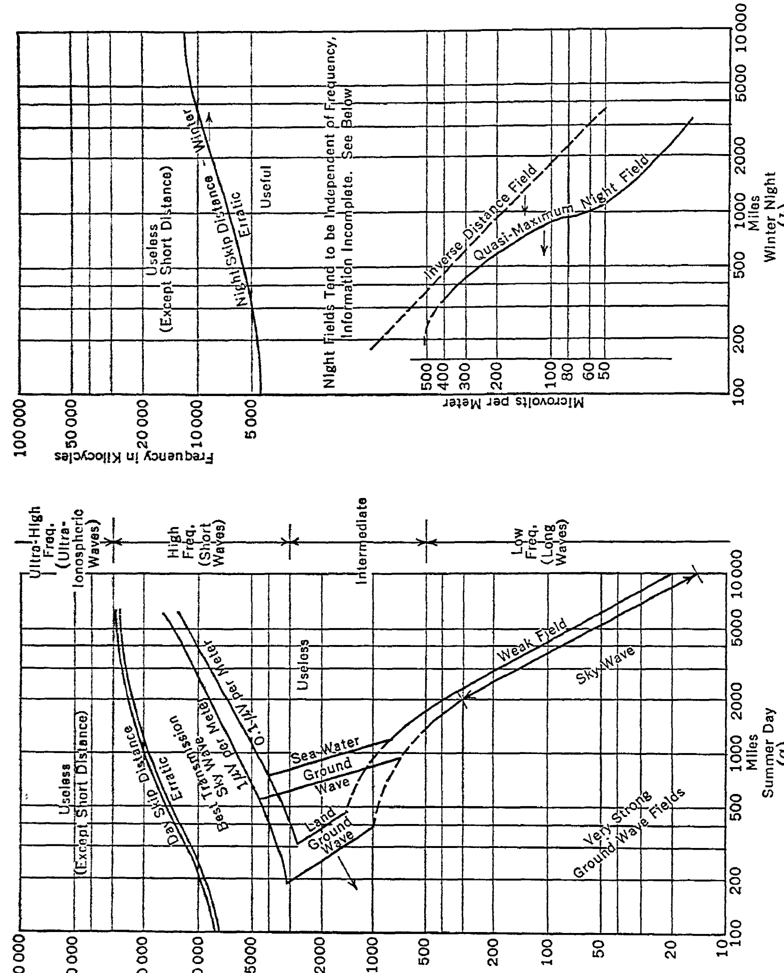


FIG. 2. General View of the Frequency-distance Relations throughout the Radio Spectrum (Vertical antennas, 1 Kw radiated)

almost certainly fails where rotation of the plane of polarization (indicating a magnetic effect) is observed. Even in the last case, the *averages* of field (as opposed to instantaneous values) usually appear to be reciprocal, and this possibly is always true at the higher frequencies.

POLARIZATION. For frequencies below 2000 kc vertical antennas are almost universally used. This is primarily because it is usually desirable to radiate with maximum efficiency in a nearly horizontal direction, which is relatively easy with vertical antennas but is impossible with horizontal antennas unless they are several wavelengths above the ground. Hence, to keep antenna dimensions within reasonable limits, horizontal electric fields are not used except for short waves.

GENERAL VIEW OF THE SPECTRUM. With respect to frequency, daylight propagation falls into natural divisions. These may be listed as follows: (1) low frequency, long distances; (2) intermediate frequency, short and intermediate distances; (3) high frequency, all distances; (4) ultra-ionospheric frequency, short distances. Here we arbitrarily describe a short distance as one from zero to 100 miles, an intermediate distance as one from 100 to 1000 miles, and a long distance as one greater than 1000 miles. Likewise, we usually think of a low frequency as one less than perhaps 500 kc. Physically, the characteristic that distinguishes a low frequency from a high frequency (3000 to 30,000 kc) is the low resistivity of the reflecting layer for the low frequencies. In fact, the ionosphere resembles a fair metallic reflector for waves of low frequencies. The wave of high frequency, on the other hand, sees in the layers of the upper atmosphere something like a reflecting plane of dielectric; the type of reflection which is most common is similar in many ways to the familiar optical phenomenon of total internal reflection. The intermediate frequencies form a transition range (500 to 3000 kc), which includes much ground-wave transmission over short and intermediate distances. Frequencies commonly designated as "ultra-high," but which are better called "ultra-ionospheric" because they are not reflected by the ionosphere, are those greater than about 30,000 kc. At night the differences between ionospheric waves of different frequencies are much less marked than by day. Waves having frequencies in excess of 1000 megacycles, more or less, are frequently called microwaves. See also Section 1, article 24.

Figure 2 gives a typical overall view of the whole radio spectrum for distances from 100 to 10,000 miles and for vertical antennas. Lines are drawn indicating the distance at which a radiated power of 1 kw would produce certain specified field strengths, e.g., 1 μv per meter. Diagram *a* represents transmission conditions on a summer day; diagram *b*, those on a winter night. On winter days, the sky wave becomes generally stronger than on summer days. As a general rule, for a given distance the highest frequency that can be received by day (the skip frequency) is greater in winter than in summer; and the lowest (absorption limit) is lower in winter than in summer. At night lower frequencies are required in winter than in summer. Transmission in the high-frequency range is markedly affected by the changes accompanying the cycle of solar activity (e.g., sunspots); since the actual phenomena are too variable to be represented by so simple a chart as Fig. 2, the comprehensive data and predictions issued by the National Bureau of Standards (Central Radio Propagation Laboratory) may be consulted to advantage in cases of actual use.

21. THE GROUND WAVE

FREE-SPACE TRANSMISSION. Historically the propagation of the ground wave has been studied by examining idealized situations. The simplest of these is the field set up in free space by a simple doublet antenna. Equations (1) and (2) give the appropriate solution and apply accurately provided that the earth is known to be without effect and the air to be a uniform and lossless dielectric. These assumptions are not true in general.

PROPAGATION OVER PERFECTLY CONDUCTING PLANE EARTH. Some situations are taken care of by assuming the earth to be a homogeneous plane and then applying the standard principles of optics. This is particularly simple if the earth in effect has *infinite conductivity*. The solution is then merely the combination of a direct wave with a reflected wave virtually coming from the "image" of the antenna in the earth plane. With low antennas and infinite earth conductivity we are led to a simple and important relation pointed out at an early date by M. Abraham. For distances short enough not to violate the assumptions as to the effective flatness of the earth and negligible attenuation, a short vertical grounded antenna of effective height h produces a field strength $\mathcal{E}_0$ in the region about it equal to

$$\mathcal{E}_0 = 120\pi \frac{hI}{d\lambda} \cos\theta \tag{3}$$

units as in (1). The formula corresponding to (2) is

$$\mathcal{E}_0 = \frac{\sqrt{90P}}{d} \cos\theta \tag{4}$$

The doubling of the numerical factor in passing from (1) to (3) merely expresses the fact that the field may be regarded as the sum of one field received directly and another by reflection from the image. Note, however, that (2) and (4) being expressed in terms of power instead of current have factors in the ratio of 1 to $\sqrt{2}$. Formulas (1) to (4) apply for distances greater than a few wavelengths. When the distance is of the order of 1 wavelength or less, the term due to *acceleration* of charge is supplemented by terms due to *velocity* and *position* (*proximity*) of charge, but for most practical purposes these may be neglected for distances greater than 1 wavelength. Except in the immediate vicinity of the antenna, the velocity of phase propagation is $2.998 \cdot 10^{10}$ cm per sec, the velocity of light in air.

PROPAGATION ALONG A PERFECTLY CONDUCTING SPHERICAL EARTH. A next step beyond that represented by (4) is the propagation of a field from a vertical grounded antenna, located at the surface of a perfectly conducting earth which is spherical rather than plane. Watson's solution [*Proc. Roy. Soc.* (A) Vol. 95, 83, 546 (1919)] is in the form of a series in which all but one term may be neglected at the greater distances. This term is as follows:

$$\varepsilon = 0.1136\varepsilon_0 \frac{d^{1/2}}{\lambda^{1/6}} \varepsilon^{-0.00376d/\lambda^{1/3}} \tag{5}$$

where ε_0 is the inverse distance field as given by (3) or (4), and d and λ are in kilometers, d being measured along the surface. This formula holds when $d/\lambda^{1/3} > 160$. For smaller distances, (3) or (4) applies. In (5), atmospheric refraction has been neglected.

The following simple empirical formula based on Watson's results and on calculations of other terms in his series carried out by C. R. Burrows can be used for distances less than about 5000 km:

$$\varepsilon = \varepsilon_0(1 + 2z)^{1/2}\varepsilon^{-z} \tag{6}$$

where $z = 0.0035\ d/\lambda^{1/3}$ and d and λ are in kilometers. All these diffraction calculations assume that the density of the lower atmosphere is independent of height; atmospheric refraction is neglected. When the deviation of refractive index from that at ground level is a simple linear function of height, refraction has the same effect as increasing the size of the earth to a virtual radius which for average conditions is about $^4/_3$ the actual radius. With refraction we have $z = 0.0029d/\lambda^{1/3}$. It is not obvious which of these values should be used. Although the refraction effect is important at sea level, it must become small at heights of several miles. Perhaps the effect can be neglected at low frequencies. At higher frequencies experiments clearly show that it should be taken into account.

TRANSMISSION ALONG AN IMPERFECTLY CONDUCTING PLANE EARTH. At distances shorter than those which give appreciable attenuation due to the shadow of the bulge of the earth, strong attenuation due to energy dissipation in the ground is found in many cases of importance in practice. Since the shadow effect is then unimportant, the solution of wave propagation over a plane of finite conductivity becomes applicable. The basic solution of this problem, due to A. Sommerfeld [*Ann. der Physik,* (4) Vol. 28, 665 (1909)], has been extended by various investigators. Owing to finite earth conductivity the wave ceases to have the simple form contemplated in connection with eq. (3) for infinite conductivity, the inverse-distance attenuation of which represents a spherically expanding wave. The field strength at the surface decreases in intensity more rapidly than the inverse of distance, owing to the absorption of energy by the currents set up in the imperfectly conducting earth. For the lower frequencies, an approximation due to van der Pol [*Jahrbuch der Drahtloson,* Vol. 37, No. 4, 152 (1931)] is useful. This is:

$$\varepsilon = \varepsilon_0 \cdot \frac{(2 + 0.3\rho)}{(2 + \rho + 0.6\rho^2)} \tag{7}$$

where ε_0 is the inverse-distance field given by (3) or (4) and

$$\rho = \frac{\pi \cdot 10^{-15}}{6} \cdot \frac{d}{\sigma\lambda^2}$$

σ is the conductivity in emu units and varies from 1 to 4×10^{-11} for sea water to 10^{-15} for very broken land; d and λ are expressed in kilometers; ρ is the "numerical distance" of Sommerfeld. Note that according to (7) the field varies inversely as the first power of distance near the transmitter and inversely as the square of the distance for $\rho \gg 20$. Also note that, since λ and σ enter only as the product $\sigma\lambda^2$, the field strength remains unaltered when the wavelength is decreased by a factor, provided that the conductivity is simultaneously increased by the square of that factor. Equation (7) holds only so long as the dielectric currents in the earth remain negligible compared with the conduction currents. This is insured if the frequency in kilocycles is very low compared with $1.8 \times 10^{-18}\ \sigma/K$, K being the dielectric constant of the ground. Numerically, the frequency should be considerably lower, for sea water, than 10^6 kc; for land, than 10^4 kc; and for fresh water, than 250 kc, the figure depending on the ground constants, which differ from place to place and to some extent with temperature.

An interesting characteristic of waves traveling along an imperfect conductor, first discussed by J. Zenneck [*Ann. der Physik,* (4) Vol. 23, 846 (1907)], is that parallel to the surface there is a longitudinal component of electric field, that is, one extending in the direction of propagation. Its amplitude and phase depend on the resistivity and dielectric constant of the ground, the phenomenon being very different over fresh water, sea water, moist ground, and dry ground. Its amplitude is zero for perfect conductivity and increases as the conductivity decreases. The phase of this component in general differs from that of the vertical component so that in the vertical plane the wave exhibits elliptical polarization. By Poynting's theorem the existence of this component is a necessary accompaniment of energy loss in the ground. The phenomenon is of importance in the design of wave antennas, which depend for their effectiveness entirely on this horizontal component. [See "Wireless Waves at the Earth's Surface" by G. W. O. Howe, *Wireless Engineer*, Vol. 17, 385 (September 1940).]

IMPERFECTLY CONDUCTING SPHERE. A still closer approximation to actual conditions is afforded by the assumptions that geometrically the earth is a perfect sphere (i.e., that local irregularities may be ignored) and that electrically over any given path it has uniform conductivity and specific inductive capacity. As to refraction in the atmosphere, the assumption made above in connection with a perfectly conducting sphere is repeated. Though these assumptions still describe a somewhat idealized picture, they nevertheless represent a large step in the direction of realism, a difficult and important mathematical task which has been successfully accomplished by the combined labors of several investigators, including Balth van der Pol. Excellent summaries, including bibliographies, have been given by C. R. Burrows and M. C. Gray [*Proc. I.R.E.*, Vol. 29, No. 1, 16 (January 1941)] and by K. A. Norton [*Proc. I.R.E.*, Vol. 29, No. 12, 623 (December 1941)]. Figures 3, 4, and 5 are based on the paper of Burrows and Gray.

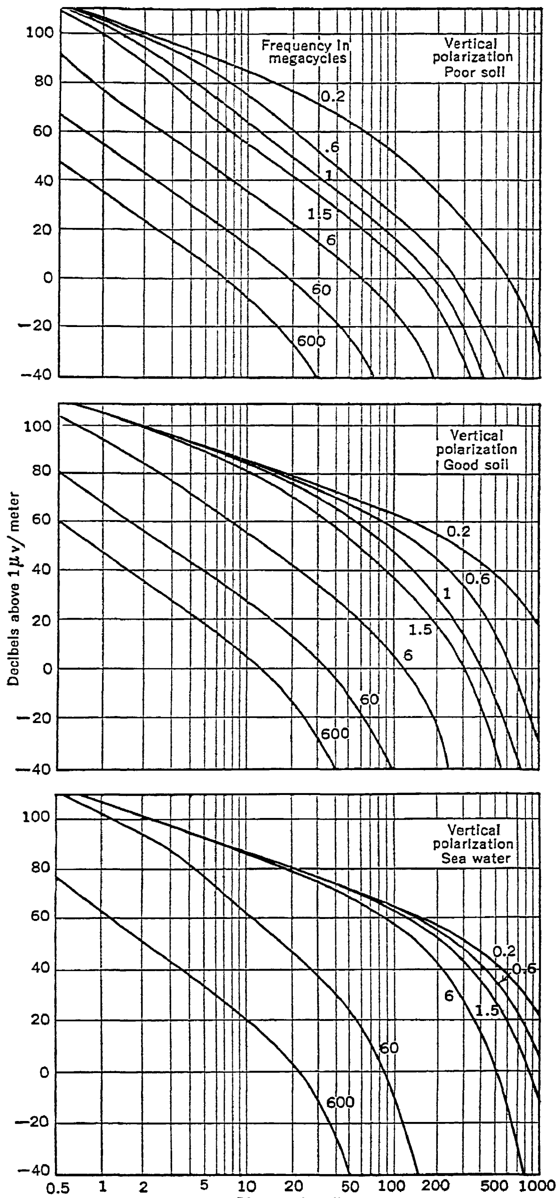


FIG. 3. Field Along Imperfectly Conducting Spherical Earth, with 4/3 Earth-radius. Short vertical antenna at ground level, measurement at ground level, for sea-water $\sigma = 4 \times 10^{-11}$ emu, $\epsilon = 80$; good soil $\sigma = 2 \times 10^{-13}$ emu, $\epsilon = 30$; poor soil $\sigma = 10^{-14}$ emu, $\epsilon = 4$, 1 kw radiated (Courtesy *Proc. I.R.E.*)

Figure 3 gives theoretical field strengths between two points on the surface of the earth when vertical polarization is used. "Standard refraction" conditions ($^4/_3$ earth radius) are assumed to occur over a sufficient depth of atmosphere to produce bending at all frequencies, a postulate frequently made which might, however, properly be questioned with reference to the lower frequencies shown, since with them transmission to the longer distances involves at intermediate points regions which are far above the earth's surface and at which the rate of decrease of refractive index with height is no longer comparable with that at the surface. The legend of the figure gives the soil constants assumed in the calculation. Important features of the curve are the "inverse-distance" tendency at the upper left corners as in the approximation of eq. (4), the rapid drop due to failing diffraction at the lower right as suggested by eq. (5), and the inverse square of distance in between as indicated by eq. (7) (for $\rho \gg 20$). A useful series of calculations along these general lines (K. A. Norton) is included in "Standards of Good Engineering Practice Concerning Standard Broadcast Stations," issued by the Federal Communications Commission, for sale by the Superintendent of Documents, Washington 25, D. C. This document also gives comprehensive information of the ground conductivities pertinent in this frequency range in the form of a United States map.

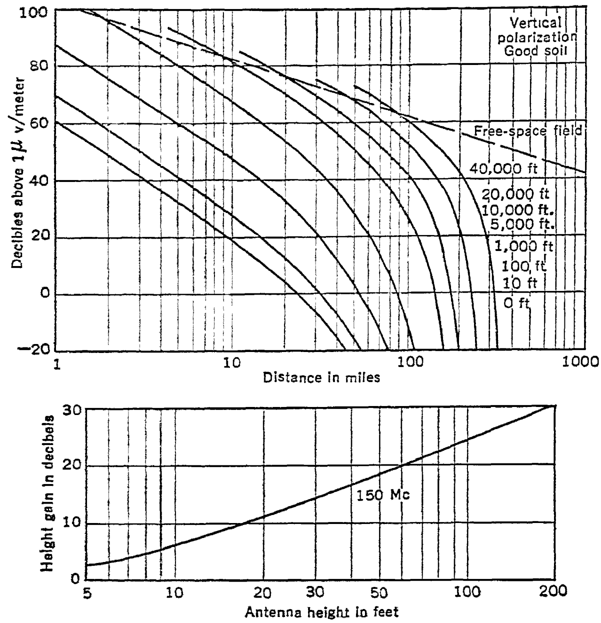


FIG. 4. Fields at 150 Mc as a Function of Height of Transmitter and Receiver, Good Soil (Courtesy *Proc. I.R.E.*)

Figure 4 illustrates propagation from one point on or near the ground to another at heights from zero to 40,000 ft above it for the special case of 150 megacycles. If the transmitting antenna (vertical polarization) is on the ground, the upper graph gives the calculated field directly. The lower graph gives the correction to be added if the antenna is elevated not more than 200 ft. The increase which accompanies the elevation of the receiver, and the inevitable failure of diffraction at greater distances, are striking features of the graph.

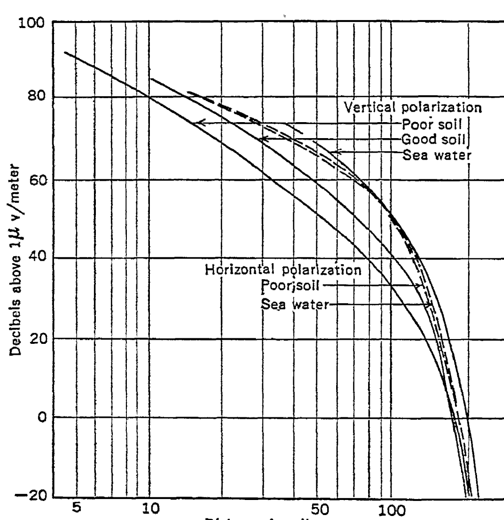


FIG. 5. Fields at 150 Mc and 10,000 Ft. as a Function of Ground and Polarization (Courtesy *Proc. I.R.E.*)

Figure 5 exemplifies for fixed antenna heights and fixed frequency how the field depends on the underlying ground and on the polarization. Note the insensitiveness of horizontal polarization to ground constants.

ULTRA-IONOSPHERIC RANGE. This is the range of frequencies higher than those capable of reflection by the ionized upper atmosphere. The dividing frequency is not at all sharp or constant but is of the order of 30 megacycles. Ultra-ionospheric waves whose length is short enough for the practical application of quasi-optical techniques such as the use of parabolic reflectors and lenses have been called "microwaves," and here an equally hazy division might be placed somewhere near 1000 megacycles.

Although the absence of reflection of ultra-ionospheric waves might hypothetically be accounted for by their penetration into a region where they are dissipated by absorption before being freed by reflection, three experiments now indicate actual passage through the ionosphere into interstellar space. In chronological order these are the detection of galactic noise by K. Jansky (and the subsequent mapping of the Milky Way by Grote Reber), the measurement of thermal radiation from the sun by G. C. Southworth [*J. Franklin Inst.*, Vol. 239, No. 4, 285 (April 1945)], and the dramatic "detection and ranging" of the moon by radar reported by DeWitt et al. [J. Mofenson, "Radar Echoes from the Moon," *Electronics*, Vol. 19, No. 4, 92–98 (April 1946)]. These experiments strongly suggest inadequacy of ionization as the reason for absence of reflections.

Qualitatively propagation in this range strikingly resembles the familiar phenomena of light. Radio "vision" tends to be limited to the optical line of sight, though *diffraction* actually extends coverage considerably beyond obstructions, such as hills or the bulge of the earth, except for extremely high radio frequencies. The earth (land or water) for many purposes may be regarded as an example of Lloyd's mirror, the ground-located receiver tending to be in the first dark fringe produced by *reflection*. *Refraction* tends to make the distant station "visible," just as it reveals the sun a few minutes before sunrise, and it produces variations and anomalies which correspond to the twinkling of stars and to the mirage. The comparison might be extended to other phenomena.

The effect of regular reflection is most pronounced in the meter range, though even in the centimeter range cleared land or water may make a good "mirror" for glancing incidence. In such cases reflection may be calculated as to amplitude and phase by means of standard optical formulas provided that an equivalent conductivity and dielectric constant are known. [P. O. Pederson, *The Propagation of Radio Waves*, Copenhagen, 1927; C. B. Feldman, "The Optical Behavior of the Ground for Short Radio Waves," *Proc. I.R.E.*, Vol. 21, No. 6, 764 (June 1933); Barfield, *J. I.E.E.*, Vol. 75, No. 452, 214 (1934); R. L. Smith-Rose, *J. I.E.E.*, Vol. 75, No. 452, 221 (1934).] As has already been implied, though reflection has a favorable effect with long waves of vertical polarization, causing the factor of 2 which differentiates eq. (3) from (1), in this range it is commonly unfavorable owing to the phase of the reflection coefficient caused by the predominance of dielectric currents in the ground at these ultra-high frequencies. The effect of reflection need not be unfavorable, however, if one or both of the terminals is located at a sufficient altitude. At high enough altitudes a regular succession of maxima and minima is encountered, whose position and amplitudes may be calculated from the amplitude and phase of reflection.

In the microwave range caution is necessary in applying the concept of reflection. Here one may not always regard the earth as the smooth and abrupt boundary between materials described simply by their conductivities and specific inductive capacities. Reflection is not always specular but is very likely to be diffuse, as the success of radar mapping proves. One should not ignore even here, however, the strong tendency toward specular reflection that glancing incidence imparts to the scattering from a rough surface.

Simple plane-wave reflection theory leads to a useful relationship which has been observed to hold with fair consistency in the range from 3 to 10 meters and even in the microwave region if the warning in the last paragraph does not apply. Over level land or over fresh water with vertical or horizontal antennas, and over sea water with horizontal antennas, the received field is:

$$\mathcal{E} = 12\pi\sqrt{5}\,\frac{\sqrt{P}}{d^2}\,\frac{H_1H_2}{\lambda} \quad \text{volts per meter} \tag{8}$$

the radiated power P being in watts, and d, H_1, H_2, and λ in meters. H_1 and H_2 are altitudes above the general reflecting area, be it ocean, valley floor, or plain. This equation is obtained by multiplying the free-space field of eq. (2) by $4\pi H_1H_2/\lambda d$, a procedure justified if $H_1H_2/\lambda d$ is less than 0.1 provided that the reflecting coefficient of the ground is near unity for the grazing incidence involved. The usefulness of eq. (8) is not limited to transmission over plane earth, as assumed in its derivation, but roughly applies also over spherical earth for transmission below the line of slight when 3 meters $< \lambda <$ 10 meters, 3 meters $< H <$ 25 meters, 1 km $< d <$ 50 km. In this extended range (8) is to be regarded as an empirical formula whose range of validity has not yet been determined. When terminals are located on hills, with a level valley between, this formula needs a correction factor due to ground reflections local to the terminals.

An interesting extension of eq. (8), valid within the same limitations, is obtained analogously to eq. (13) in the section on radio antennas. If G_T and G_R are the power gains of the antennas (in terms of an "isotropic radiator"), and P_T and P_R the powers transmitted and received, it can be shown that

$$\frac{P_R}{P_T} = G_T G_R \left(\frac{H_1 H_2}{d^2}\right)^2 \tag{8a}$$

Note that power received is independent of frequency. If the antennas are short doublets, $G = 1.5$. (K. Bullington, *Proc. I.R.E.*, 1947.)

With vertical antennas over sea water, but otherwise with conditions specified above, an approximate inverse-square-of-distance variation has been found up to 30 or 40 km, with the important difference that propagation became poorer as the frequency was raised. This is in accord with simple optical theory, due account being taken of the electrical constants of sea water. Equation (8) therefore does not apply here, though it does for horizontal antennas over sea water.

Since diffraction extends the range beyond obstacles it is a favorable factor. Ignoring diffraction and refraction we could at most transmit only to points above the line of sight. The range would then be limited to $D = 3500(\sqrt{H_1} + \sqrt{H_2})$, all distances being in meters. Within this range eq. (8) is applicable unless the antennas are too high, with qualifications already mentioned. Note that line of sight does not at all guarantee the free-space field strength.

It is possible to make instructive calculations of the fields behind obstructions, such as hills, by application of the standard mathematical theory of diffraction. Thus, the behavior of a knife edge in the familiar optical example becomes a guide in radio transmission past obstacles. In a radio problem, in order to obtain a reasonable estimate of this kind, the effect of ground reflection at the transmitter and at the receiver needs to be taken into account.

If the index of refraction of air is calculated as a function of height from average meteorological conditions [Humphreys, *Physics of the Air*, McGraw-Hill (1940), p. 80], the gradual decrease of index leads to a "standard" or "normal" condition which can be taken into account in a simple manner. Such calculations and theoretical considerations indicate that, if the topographic cross-section of the path is plotted as though the earth had a radius $^4/_3$ times its actual radius, the solution of the corresponding problem assuming a uniform refractive index is also the solution for the actual problem including the effect of refraction. We have already had occasion to use this method above in the "spherical earth" problem. In dealing with microwaves, however, this simplication will be misleading if the possibility of many other distributions of refractive index is forgotten.

In connection with the foregoing topics, reference may be made to articles in *Proc. I.R.E.*, Vol. 21, No. 3 (March 1933), by Jones (p. 349); Trevor and Carter (p. 387); Schelleng, Burrows, and Ferrell (p. 427); and Englund, Crawford, and Mumford (p. 464); also Englund, Crawford, and Mumford, *B.S.T.J.*, Vol. 14, No. 3, 369 (July 1935).

At shorter wavelengths refractive variations within a small range of height may become important because the wave in traveling between two points occupies only a small fraction of a kilometer (that is, the necessary Fresnel zones are now included in a small transverse area). It results, for example, that such waves may be trapped beneath a level of minimum refractive index and may travel unusually long distances, and that under other conditions they may unexpectedly fail over short ones.

Refraction theory as applied to extremely short waves has led to the use of a *modified refractive index* of the air, M, as a function of height, h. If the actual index as ordinarily used is $n(h)$ and the radius of the earth is a, the definition $M \cdot 10^{-6} = n(h) - 1 + h/a$ leads to the same solution with an assumed flat earth that the actual index leads to with curved earth, the scale factor 10^{-6} being used for numerical convenience.

Just as short waves are "bent down" by the ionosphere because its refractive index decreases with height—that is, the phase velocity increases with height—so in the troposphere a decrease in modified index will tend to confine microwaves beneath it. Indeed, if meteorological conditions are such as to give a maximum index at a certain level with progressively smaller values above and below, one would on ray theory expect the ray in its horizontal passage to undergo consecutive upward and downward bendings about the level of maximum index (minimum velocity). Similarly the wave might be confined between an index which decreases with height and the reflecting floor of the ocean (or perhaps of land). Such phenomena do occur, and their importance is that they may lead to abnormally high or abnormally low fields. The reason for the strong fields is that cylindrical rather than spherical expansion causes a slower decrease with distance, i.e., increased range horizontally. The phenomena resulting are likely to be complicated and variable, though on the other hand such "anomalies" may be so consistent as to be the normal condition. ["9 cm and 3 cm Propagation in Low Ocean Ducts" by M. Katzin, R. W. Bauchman, and Wm. Binnian, Dept. of Com., Office of the Publication Board, Report PB 13747 (1945); "Wave Theoretical Interpretation of Propagation in Low Level Ocean Ducts" by C. L. Pekeris, Dept. of Com., Office of the Publication Board, Report PB 20228.] Another way of looking at these ducts is to use wave guide concepts: in fact, a duct is a wave guide with many modes of propagation which are excited to different extents depending on the elevation of the transmitter with reference to that of the duct. The second article cited discusses the subject from that point of view. For a summary of the war work dealing with the various aspects of this problem see *Radio Wave Propagation, Consolidated Summary Technical Report of the Committee on Propagation*, by Burrows and Attwood, Academic Press, New York.

With wavelengths longer than 10 cm the absorption of energy from the wave due to the atmosphere itself is not important for practical purposes, but with shorter waves at least three known mechanisms may have to be considered. *Water in the liquid* or *solid* phase (e.g., rainfall) is one of these that can become very serious at wavelengths of a centimeter or two. The loss depends on the total water per unit volume and on the size of drop or particle and is due to scattering. Although this is harmful to radio transmission, the phenomenon is being utilized by meteorologists in the detection and location of storm areas by means of radar. In the *vapor phase, water* has its longest wave resonance at 1.3 cm, so that near this wavelength high absorption is to be expected and is found for high humidities. *Oxygen* has its first resonance absorption at 0.5 cm.

Directional properties of tropospheric waves have been studied at a wavelength 3.25 cm by W. M. Sharpless, and at 1.25 cm by A. B. Crawford and W. M. Sharpless, *Proc. I.R.E.*, Vol. 34, No. 11, 837–848. Deviations in azimuth were found to be at most of the order of 0.1°. Although in elevation the deviations were several times as great as this, they were never large. Fading about the free-space field strength was observed, and at times the field exceeded this by a factor of 4 (12 db). At times there were multiple waves coming in from slightly different directions, the variations of which produced fading.

22. THE SKY WAVE

THE IONOSPHERE. Whereas for most radio services up to a few hundred miles transmission depends on the direct ground wave, for all long distances successful transmission depends on the existence of a "ceiling" in the upper atmosphere which, by returning to the earth the outgoing waves, lays down a useful signal at distant points where the ground wave is negligible. Even the longest waves used in radio communication depend on such "sky waves" for distances beyond a thousand miles or two. In this section we shall describe present views of this ceiling, the common name for which is the "ionosphere," formerly called the "Kennelly-Heaviside layer." The first suggestion that an electrically conducting region exists in the upper atmosphere was made by Balfour Stewart to explain variations in the magnetic field of the earth. Kennelly and Heaviside were the first to see the necessity of such a region for explaining radio transmission phenomena.

At sea level, the atmosphere is scarcely conducting at all, but as the elevation is increased the conductivity increases owing to an increase in the number of ions in unit volume. This ionic density increases both because the major sources of ionization are outside the earth and because, at the lower pressures encountered, the ions last longer before recombination neutralizes them electrically. Heavy ions (e.g., ionized molecules) have much less effect on radio waves than electrons. Electrons are known to exist in sufficient quantity in the upper atmosphere to produce most of the effect observed, but the radio effects of heavy ions have not as yet been definitely identified. The number and distribution of these ions depend on various factors, including altitude, geographical and geomagnetic latitude, local time, time of year, and solar activity (e.g., sunspots). The variation with altitude is very important. It has been found that the increase with altitude is not at all uniform and simple, but that there are regions where the density attains, or tends to attain, maximum values. Although methods thus far devised are not suited for studying those levels above the maxima where the density may actually decrease with height, it is probable that one or more of such decreases actually occur. These levels of maximum (or tendencies thereto), frequently called "layers," are illustrated in Fig. 6. The maximum occurring somewhat above 100 km is known as the E layer, and those above 160 km as the F layers, names which will perhaps be superseded when the mechanism of their production is explained. It is fairly definitely known that ultraviolet light from the sun is an important source of ionization, particularly in the E region and in one region of the F layer. This leads to the large differences between the day and night behavior of radio waves, in particular to phenomena occurring at sunrise and sunset.

Another cause of ionization is particles from the sun, which most authorities regard as the radiation by which solar disturbances communicate the major disturbing effects to the earth. These are in particular thought to be the cause of the absorbing "clouds" which seem to form in polar regions beneath the E layer and hinder short-wave transmission during magnetic storms. The distribution of ionization changes with time, and Fig. 6 is to be regarded merely as typical. Tendencies toward maxima come and go, and both the E and F regions seem to be composites of two or more layers. The E layer is the most consistent and is the most important one in the broadcast band (1500 kc and lower) and at lower frequencies. For these lower frequencies, the waves do not ordinarily penetrate higher than the E layer. For the lowest frequencies it is not known whether reflection occurs at the E layer or at some lower level, there being some evidence that the height is 80 or 90 km. The F region is of most importance for the short waves, particularly over longer distances, but the E region also contributes components to the signal, which is usually very complex.

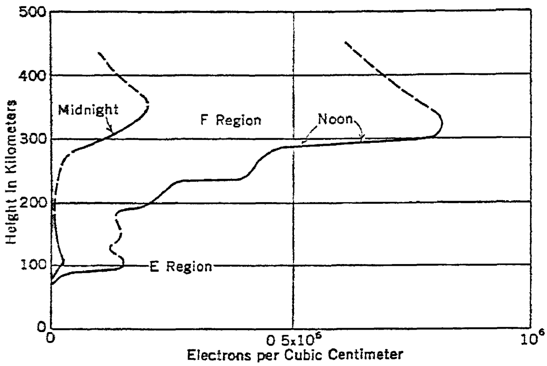


FIG. 6. Concentration of Electrons in the Upper Atmosphere. The D region is the ionosphere below 90 km, the F region is the part above 160 km, and the E region is the part between these limits.

The most fruitful method of studying these regions has been the pulse or echo method of Breit and Tuve [*Phys. Rev.*, (II) Vol. 28, 554 (1926)], in which a very short-wave train of a given frequency is radiated upwards and the times required for various reflections to be returned are obtained with an oscillograph, preferably of the cathode-ray type. These times may be converted into virtual heights of the reflecting layer by assuming the pulse to travel with the velocity of light. This method of "radio detection and ranging" is in fact one of the forerunners of radar. Actually, the pulse travels with a slower group velocity than this while in the ionized region, but as a result of making measurements at several frequencies it is often found that the virtual height within limits is nearly independent of frequency, and for such frequencies the actual cannot be very different from the virtual height. Transmission through the ionized region is complicated by the earth's magnetic field, which makes of it a doubly refracting medium in which the wave is broken into two components of different polarization traveling with different velocities. It is this characteristic which has led to the identification of the electron as the active ion. The magnetic field leads to complications in practical communication by causing rotations in the plane of polarization, leading to one type of fading (see later section on fading) and to errors in direction finding with loop aerials.

It is natural to suppose that the electrical characteristics of the ionized region are linear, so that different disturbances may be superposed without interaction. Evidence has been found, however, that this is not invariably true [B. D. H. Tellegen, *Nature*, p. 840 (June 10, 1933)]. If two broadcasting stations of high power operate on entirely different wavelengths and are separated by some hundred kilometers, modulation originally impressed on one has been found under certain conditions to have been transferred to the wave of the second. This indicates that the ionosphere does not have strictly linear characteristics. It is called the Luxemburg effect.

For detailed information on the ionosphere, and for bibliographic references, the following may be of interest: E. V. Appleton, *Inst. E.E. (London)*, Vol. 7 (September 1932); Kirby, Berkner, and Stuart, *Proc. I.R.E.*, Vol. 22, No. 4, 481 (April 1934); Schafer and Goodall, *Nature*, June 3 and Sept. 30, 1933; Dellinger, *Trans. A.I.E.E.*, *Supplement*, Vol. 58, 803 (1939); Darrow, *Bell Sys. Tech. J.*, Vol. 19, No. 3, 455 (July 1940).

SKY-WAVE PROPAGATION. It is the general belief that waves which travel long distances do so by means of multiple reflections, although the suggestion has been made that short waves (e.g., frequencies above 3000 kc) do so in a single step. Thus, in Fig. 7 a wave is conceived to travel from A to B by three reflections from the E region (3), or by two from the F region (2). The single-step path is represented by curve 1. If at each ionospheric reflection double refraction due to the earth's magnetic field were to occur, for the two-reflection wave not one but four (2^2) components might be found, and for the three-reflection mechanism, eight. Whether the complexity is thus explained or not, it is a fact that the received wave is frequently very complicated. As Fig. 7 suggests, there is wide diversity among components in their angles of elevation. In general, however, for all these paths, the received energy arrives with a downward component of velocity. This is of great importance in practice, since the mode of transmission places important directional requirements on the antennas at the two terminals. In the horizontal plane, radio waves as a general thing travel along the great circle defined by the locations of the terminals. Consequently, the waves on arrival are usually directed approximately along

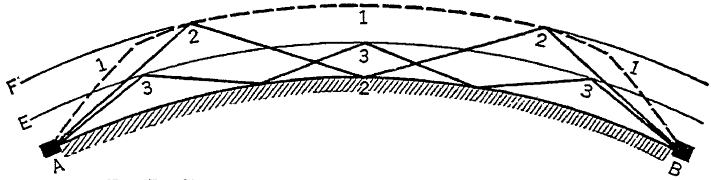


FIG. 7. Sky-wave Propagation according to the Ray Theory

the true bearing of the transmitter, regardless of frequency range. Small variations and differences exist in azimuth, with occasional large ones. See Friis, Feldman, and Sharpless, *Proc. I.R.E.*, Vol. 22, No. 1, 47 (January 1934); Friis and Feldman, *Proc. I.R.E.*, Vol. 25, No. 7, 841 (July 1937).

LOW FREQUENCIES (LONG WAVES). Among the chief characteristics here are: (1) at a given frequency, the diurnal variation of the field and the difference between day and night attenuation; (2) by day, the greater attenuation at the higher frequencies; (3) by night, the relatively low attenuation and the relative independence of attenuation on frequency; (4) seasonal variations; (5) propagation substantially along the great-circle path and the departure and arrival in a substantially horizontal direction; and (6) the practical absence of fading.

Typical diurnal variations of field strength from American transatlantic long-wave stations as received in England are shown in Fig. 8, which is reproduced from Espenschied, Anderson, and Bailey, *Proc. I.R.E.*, Vol. 14, No. 1, 7 (February 1926). The times when the path is entirely in daylight, entirely in darkness, and partially in each are shown by the shading in the strip at the bottom.

Both seasonal variations and diurnal variations are brought about by the changing position of the path relative to the hemisphere illuminated by the sun. In the summer, the duration of the daylight transmission phenomena is naturally longer than in the winter on account of the longer days. It is this change in the lengths of the day and night periods which is the most striking feature of the seasonal variation, rather than any change in the strength of signal. When the entire path is illuminated by the sun, or when the entire path is in darkness, the characteristic day or night phenomena are observed. When the path is half illuminated, half darkened, a characteristic minimum may be found in the diurnal curve. This is illustrated in Fig. 8, which shows a pronounced minimum occurring near sunset in the 57,000-cycle curve.

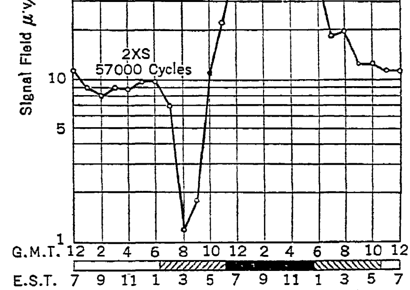


FIG. 8. Diurnal Characteristics of Low Frequencies in Transatlantic Propagation

It is evident that the field to be obtained at any time cannot be predicted by any simple formula, but if precision is not required it is possible to determine its order of magnitude for the night and for the day condition. Very roughly, the midnight field in long-distance transmission has an average of the order one-fifth the inverse-distance value. By day, the fields are more consistent and the average values are indicated by the Austin-Cohen formula,

$$\varepsilon = \varepsilon_0 \sqrt{\frac{\theta}{\sin\theta}}\, \varepsilon^{-46 \times 10^{-6} D f^{0.6}} \tag{9}$$

where ε_0 is given by (3) or (4), θ in radians is the angle subtended at the center of earth by the path, D is in kilometers, and f is in kilocycles and is less than 1000 kc. These constants were suggested by Austin in 1926 [*Proc. I.R.E.*, Vol. 14, No. 3, 377 (June 1926)] in order to make the formula more nearly universal for daytime transmission than the original formula, in which the exponent was $-87 \times 10^{-6}Df^{0.5}$ A convenient form is

$$\varepsilon = 3 \times 10^5 \frac{\sqrt{P}}{D} \cdot \sqrt{\frac{\theta}{\sin\theta}} \cdot \varepsilon^{-46 \times 10^{-6}Df^{0.6}} \tag{10}$$

in which P is radiated power in kilowatts and E is in microvolts per meter. The difficulty of expressing transmission data in this simple and usable form is brought out by the fact that in one part of the frequency range covered, namely from 17 to 60 kc, data received in transatlantic transmission are better represented when the exponential factor is modified to $\varepsilon^{-4 \times 10^{-6}Df^{1.25}}$ [Espenschied, Anderson, and Bailey, *Proc. I.R.E.*, Vol. 14, No. 1, 7 (February 1926)].

Effects accompanying magnetic storms, and secular variation, are discussed under "Solar Disturbances" in article 23.

INTERMEDIATE FREQUENCY. As indicated in Fig. 2, the daytime field of the sky wave in this range, broadly speaking, is attenuated beyond the possibility of usefulness. Apparently enough ultraviolet light from the sun penetrates to levels of the order of 100 km to maintain an absorbing stratum of ionization in spite of relatively rapid recombination. Near sunset, however, this cloud disappears, permitting a considerable reflection to distant points on the ground during the night. There is still absorption, and it is variable as Fig. 9 shows, but propagation to long distances is ordinarily possible. (See Standards of Good Engineering Practice, loc. cit., from which Fig. 9 is reproduced.) Note that this figure is reasonably consistent with the value one-fifth of inverse-distance field mentioned in the preceding section for low frequencies.

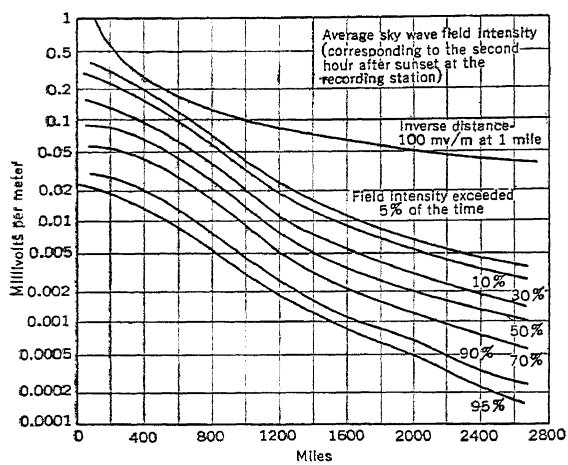


FIG. 9. Night-time Field Strengths from 250 to 2700 Miles (F.C.C. Data)

To a first approximation night-time transmission in this range is independent of frequency.

HIGH FREQUENCIES (SHORT WAVES). In contrast with the low-frequency range discussed above, there is a range of frequencies above approximately 3000 kc in which daylight sky wave transmission improves with increasing frequency, though this trend is subsequently reversed. This short-wave range is limited at the high-frequency end by the inadequacy of electrons per unit volume of the ionosphere. The limiting frequency in the daytime is not very different from 30,000 kc (10 meters). Both limits are variable, and the figures given are somewhat arbitrary. Among the chief characteristics of transmission in this range are: (1) the diurnal variation of field strength and the prominence of day-to-day fluctuations; (2) the greater distances of transmission obtained with the higher frequencies, especially by day; (3) the "skip" effect, or the existence of a region about the transmitter in which the direct wave is absent owing to attenuation of the ground wave, and the sky wave, if present at all, is weak and erratic owing to electron limitation; (4) habitual fading, sometimes of extreme rapidity, and the common occurrence of selective fading; (5) the necessity in most cases for more than one frequency for 24-hour service; (6) the great reduction of field strength in northern and southern latitudes concomitant with magnetic storms and, by contrast, the absence of a pronounced effect in equatorial regions, and other phenomena having a solar origin; (7) a secular variation following the 11-year sunspot cycle; (8) great-circle transmission and a wide variety of angles in the vertical plane.

Diurnal Variation. Typical diurnal variations are shown in Fig. 10, which depicts the changes occurring in a 24-hour interval over a path between Deal, N. J., and New Southgate, England, the radiated power being 1 kw. The curves bring out the advantages of the higher frequencies by day and of the lower by night. A typical curve on an intermediate frequency also is shown (10.55 mc). Day-to-day variations are more pronounced on such intermediate frequencies than for either higher or lower frequencies. On some days the intermediate frequency transmission may resemble the lower frequencies, on others the higher.

Variation with Distance. For a given distance, the transmission conditions depend on the geographical latitude of the stations and to some extent on the geomagnetic latitudes, the difference in longitude, the time of the day, the time of the year, and the time in the solar cycle. Figure 2 will serve as an approximate indication of the frequencies suitable for various distances for day and for night. The curves were based largely on data obtained during the years 1926 to 1930, a period which included a sunspot maximum. In general, the frequency required is higher the lower the latitude, the nearer the time to noon at the midpoint of the path, and the nearer to the secular sunspot maximum. In the winter the maximum usable (MUF) frequency for midday is greater than in summer, but the lowest usable high frequency (LUHF) is lower. In a diurnal curve, the characteristic day and night conditions are obtained for a longer period the more uniform the conditions along the path. Thus, a north-south path has a more abrupt change from day to night conditions than an east-west path, and the transition condition is of shorter duration. This transition period is relatively difficult because no single frequency can be best adapted to both the day and night portions of the path. Hence, long east-west paths tend to be more difficult than north-south paths. Usually, a day frequency gives a weaker field by day than a night frequency does by night. The variation of best frequency with secular magnetic change is exemplified by transatlantic paths such as that from New York to London. During the sunspot maximum of 1930, the best daytime frequency was about 18,000 kc, whereas, during the minimum which followed, the best was under 15,000 kc. (See later discussion under "Maximum and Minimum Usable High Frequencies.")

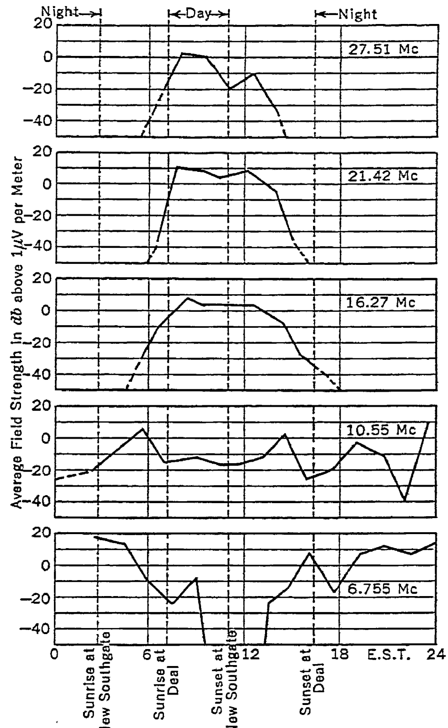


FIG. 10. Diurnal Characteristics of High Frequencies in Transatlantic Propagation

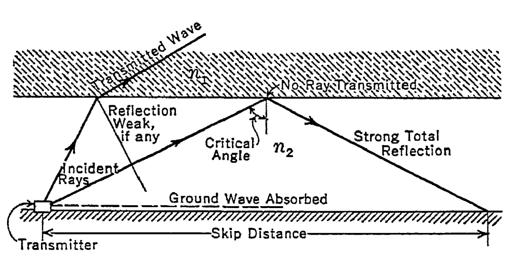


FIG. 11. The Skip Distance as a Phenomenon of Total Reflection

The Skip Effect. This effect is in many ways analogous to the phenomenon of total reflection in optics. A light wave (Fig. 11) passing from one medium of dielectric constant n_2 into another of smaller dielectric constant n_1 is subject to reflection at all angles if the change in dielectric constant is abrupt, and to *total* reflection for angles of incidence greater than $\Psi = \sin^{-1} n_1/n_2$. In radio, the type of reflection resembling total reflection is believed to be the more important, since the change in dielectric constant is gradual. For large angles of incidence (incident ray approaching the horizontal) the wave may be reflected to a great distance, but for angles less than the critical angle Ψ (closer approach to the vertical) the wave may pass through without substantial reflection. Although the radio problem is more complex than that of the optical illustration owing to the gradualness of the change in refractive index brought about by the gradual change in ionic density, in both cases the strong wave of total reflection is absent at the smaller angles of incidence. Since the index, n_2, of air at sea level is unity, waves incident at the ionosphere at an angle less than $\sin^{-1} n_1$ pass through. n_1 is to be taken as the index of the ionized region at the altitude of maximum ionic density (minimum refractive index), and, since it is a function of frequency, time, and other factors, the skip distance is a function of the same factors. Figure 2 shows approximate values of skip distances, obtained experimentally, for day and night conditions as a function of frequency. [Taylor and Hulburt, *Phys. Rev.*, Vol. 27, 189 (February 1926).]

Directions of Departure and Arrival. For a single wave component, this direction can be expressed by two angles. One is that in the horizontal plane and can be given either in terms of true north or as deviations from the great circle containing the transmitter. The other is the angle in the vertical plane and is usually given as the angle between the ray and the horizontal plane.

In the horizontal plane, the direction of arrival does not ordinarily deviate markedly from the great circle, deviations greater than a few degrees being unusual.

In the vertical plane, the angle with the horizontal may be anywhere in the range from 0° to 90°, depending on conditions in the ionosphere, distance between stations, and frequency. In general, as constructions such as that of Fig. 7 would indicate, the angle for short paths is high. For long distances, the angles tend to be small. Thus, in transatlantic communication, angles from 10° to 20° are common. Angles as low as 8° and as high as 38° have been measured. The average seems to be not far from 15°. On the other hand, signals received near New York from Buenos Aires commonly arrive at vertical angles less than 5°. In the case of the transatlantic paths, a great deal of attention has been given to these questions by Friis, Feldman, and Sharpless, *Proc. I.R.E.*, Vol. 22, No. 1, 47 (January 1934), and by Feldman, *Proc. I.R.E.*, Vol. 27, No. 10, 635 (October 1939). They found that the directions of individual wave components do not change rapidly or capriciously; that the components which arrive at the higher angles arrive later than those at the lower, qualitatively as might be expected from Fig. 7. In the vertical plane, angular spreads between lowest and highest component have been found at times to be smaller than 1° and at other times as large as 20°.

Polarization. In sky-wave transmission of high frequencies the composite polarization of the received wave is on the average independent of that transmitted. The direction of the electric field changes in a random manner with a rapidity which is connected with that of fading.

Echoes of Long Delay. Among the more unusual echoes of long delay, two types are of particular interest. Under certain conditions, "round-the-world echoes" can be observed. These are waves having a delay of about $^1/_7$ second, which travel all the way along a path which probably is not very different from the great circle separating the day and the night hemispheres. They are not usually observed but are prevalent at certain times of the year for a given pair of stations.

A second type of echo exhibits delays as great as 30 seconds. This extraordinary retardation may be due to extremely low group velocities in the ionosphere or to waves which travel long distances outside the ionosphere before they return. They are as rare as they are mysterious.

Maximum and Minimum Usable High Frequencies. Several years ago the National Bureau of Standards began a systematic study and publication of the month-to-month variations in the ionosphere at Washington, D. C., with predictions of transmission conditions to be expected. During the war this work was developed comprehensively. From the point of view of operation, it is important to know for a given time and path the *maximum usable frequency* (MUF) permitted by the skip phenomenon and the *lowest usable high frequency* (LUHF) permitted by ionospheric absorption. Predictions on a worldwide basis are available. (See publications of the Central Radio Propagation Laboratory, National Bureau of Standards, Washington 25, D. C.)

23. OBSTACLES TO TRANSMISSION

ATMOSPHERIC INTERFERENCE. "Atmospherics" or "static" are electric waves of natural origin which often mar radio reception or make it impossible. The sounds produced vary from the crackling of extremely short impulses called "clicks," such as may be produced by local lightning flashes, to the steady background roar called "grinders." Hisslike atmospherics have also been observed. The principal characteristics of atmospherics may be listed as follows:

1. They are more intense in summer than in winter, regardless of radio frequency. In the northern hemisphere they reach their maximum in July or August. This leads to more difficult transmission at that time.

2. They have a diurnal variation in intensity. For frequencies below about 10,000 kc the night-time intensity is greater than the daytime. The difference is greatest at about 1000 kc in the vicinity of New York City, being perhaps 50 db, and is very small at 15 kc. For the octave above 10,000 kc atmospherics are strongest in the daytime (see Fig. 12).

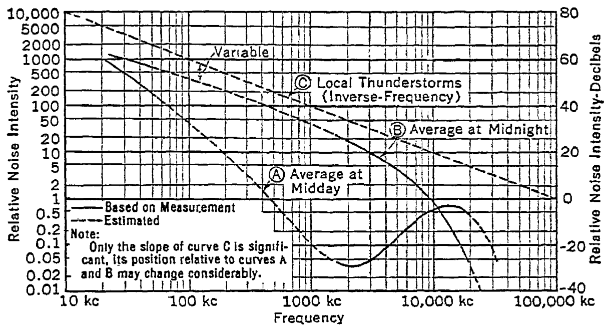


FIG. 12. Atmospheric Noise as a Function of Frequency in the Vicinity of New York City

The diurnal variation has a definite relation to the diurnal variation in the transmission characteristics of the frequency used. [C. N. Anderson, *Proc. I.R.E.*, Vol. 21, No. 10, 1462 (October 1933).]

3. The law of variation of amplitude of atmospherics with frequency cannot be simply stated and is in fact not definitely known. Some evidence favors an inverse first, and some an inverse second power. Figure 12 gives one estimate of the frequency distribution [R. K. Potter, *Proc. I.R.E.*, Vol. 20, No. 9, 1514 (September 1932)]. The figure is for reception in the northeastern part of the United States; elsewhere the intensity and its distribution with frequency would probably be somewhat different. By day the amplitude apparently follows approximately the inverse square of the frequency up to 1000 kc. At higher frequencies an increase having a maximum at about 15,000 kc sets in, but at this frequency noise originating in the receiver is likely to be more important. The night curve seems to follow the inverse first power up to about 10,000 kc. The advantage of low frequencies due to better transmission is evidently reduced by the greater amount of atmospheric disturbance encountered. Likewise, for the low frequencies, the advantage due to the fact that the fields at night are stronger than day fields is similarly reduced by the fact that atmospherics are then stronger.

4. Atmospherics are predominantly of tropical origin. Exceptions are those on intermediate frequencies (500 to 3000 kc) during the day, and atmospherics on all ultra-ionospheric waves. These are of comparatively local origin. On the low frequencies, less atmospheric interference is found the farther the receiver is removed from the tropics. For example, they are less in Maine than near New York. [Austin, *Proc. I.R.E.*, June 1926, p. 373; Espenschied, Anderson and Bailey, *Proc. I.R.E.*, Vol. 14, No. 1, 7 (February 1926).]

5. They are said to be stronger on land than over the ocean. [Austin, *Proc. I.R.E.*, Vol. 14, No. 1, 133 (February 1926).]

6. Atmospherics arrive at a receiver from all directions, but usually certain general directions predominate and over a small period of time the directivity may be comparatively sharp. This is important in reception, since directional discrimination in the receiving antenna can be used to reduce the interference without reducing the desired signal; this is not possible if the desired signal arrives from the same direction as the atmospherics. When there is a variable null direction in the antenna polar diagram, the improvement obtainable is very considerable unless the disturbances arrive through a wide range of angles. [Espenschied, Anderson, and Bailey, *Proc. I.R.E.*, Vol. 14, No. 1, 7 (February 1926).]

This directional characteristic is due to the existence of broad centers of origin which seem to coincide with thunderstorm centers, most of which are over land. Among those of greatest importance to reception in the United States are Ecuador, Brazil, and Central Africa in the winter, and Mexico and Central America and the waters between there and

Florida, Florida, and New Mexico in the summer. The sources actually causing interference will depend on the distance range of the frequency used. Thus, frequencies of the order of 1000 kc will have only local static by day, and ultra-ionospheric waves will never be troubled by distant sources. At night, all services using the sky wave may be exposed to distant sources. The very low frequencies are exposed to distant sources at all times. (See papers by Dean and by Harper, *Proc. I.R.E.*, July 1929, pp. 1185 and 1214.)

7. Atmospherics travel along the earth in the same manner as signals and are therefore subject to the same laws of attenuation. Their diurnal variation is explained principally by this fact.

8. Atmospherics apparently originate in discharges with sufficiently abrupt wave fronts to shock-excite circuits tuned to any radio frequency lower than 150 mc (or some higher limit; see next paragraph). Components also fall in the audible range with frequencies well below 1 kc. One class called "tweeks" has a limiting frequency between 1600 and 1700 cycles, suggesting multiple reflections of a pulse between the earth and a conducting layer 90 km above it. [Appleton, Watson-Watt, and Herd, *Proc. Roy. Soc.*, A, Vol. 111, 165 (1926); Burton and Boardman, *Proc. I.R.E.*, Vol. 21, No. 10, 1476 (October 1933).]

9. At 150 mc Schafer and Goodall ["Peak Field Strength of Atmospherics Due to Local Thunderstorms at 150 Megacycles," *Proc. I.R.E.*, Vol. 27, No. 3, 202–207 (March 1939)] found (*a*) that the peak intensity of disturbances varies 20 db between different storms at the same distance; (*b*) for nearby storms the inverse distance relation is a good approximation for the calculation of the variation of peak disturbance with distance; (*c*) the use of high instead of low receiving antennas increases the signal-to-disturbance ratio almost directly with height for storms within 10 miles; (*d*) the durations of some of the narrower peaks in any particular lightning discharge are as short as a few microseconds or shorter; (*e*) the maximum equivalent peak field for a storm 1 mile away was about 0.015 volt per meter with a band width of 1.5 megacycles. Although it is a common belief that atmospherics do not exist at microwaves, actual measurements do not seem to have been reported.

10. Frequencies in the band near 20,000 kc exhibit distinctly an aural difference between atmospherics from local and from distant sources. Atmospherics from local sources give the "crash" type; those from distant sources ordinarily give a fairly steady weak background. The difference is not the result of any dissimilarity in mechanism but is due to the skip effect which excludes disturbances from intermediate distances. The direction of arrival of the steady background corresponds with that observed on long waves (SW to SE near New York), and the disturbance is heard only when long-wave static is very strong. As the frequency is increased into the ultra-ionospheric range, the steady background disappears, so to speak, in the distance, leaving only the crashes due to occasional local storms. A weak hisslike disturbance apparently from a fixed direction in space and from distances beyond the confines of our solar system has been observed on 20,000 kc. [K. Jansky, *Proc. I.R.E.*, Vol. 20, No. 12, 1920 (December 1932), and Vol. 21, No. 10, 1387 (October 1933).] Jansky pointed out that this noise arrived from the direction of the galactic center (Sagittarius), and Reber [*Astrophysical Journal*, Vol. 100, 279–287 (1944)], at a considerably higher frequency which permitted high directivity, explored the region of the Milky Way and found noise contours corresponding to it. Black-body radiation from the sun has been studied by G. C. Southworth [*J. Franklin Inst.*, Vol. 239, No. 4, 285 (April 1945)] at a wavelength of 3 cm; it may be looked upon as resistance noise, the resistance being radiation resistance which has a high "temperature" when the antenna points at the sun. An enormous increase in this noise has been reported to occur at times of abnormal solar activity. [Pawsey, Payne-Scott, and McCready, *Nature*, Vol. 157, No. 3980, 158 (Feb. 9, 1946), and Hey and Stratton, *Nature*, Vol. 157, No. 3976, 47 (Jan. 12, 1946).]

11. The noise level differs in different parts of the sunspot cycle. At high frequencies there is a direct and at low an inverse relation to sunspot numbers.

12. Methods of combating atmospherics must be based on the use of some characteristic in which the wave of the signal differs from that of the disturbance. The most important of these are frequency, direction, and amplitude. In the first, selective circuits are used which suppress current of frequencies not present, or necessary, in the signal. Even with absolutely ideal selective circuits, an irreducible minimum of energy will pass through them, and this minimum increases linearly with the frequency range necessary for signaling. (Carson, *B.S.T.J.*, April 1925.) Directional discrimination has already been discussed in paragraph 6 above. As regards amplitude, the most obvious procedure is to increase the effective radiation toward the receiving terminal by increasing either the power capacity of the transmitting set or the effectiveness of the transmitting antenna. Another is the use of the "compandor" in telephony, by which the low amplitudes are raised above their natural value while passing from the transmitter (through the part of the circuit exposed to atmospherics) to the receiver, the normal values being usually restored subsequently.

An old method of discrimination against atmospherics of high amplitude uses opposed detectors, equally sensitive at high amplitudes but sufficiently different at low amplitudes so as not to cancel the signal. [Englund, *Proc. I.R.E.*, Vol. 16, No. 1, 27 (January 1928).]

With amplitude modulation (in which intelligence is carried by variations of the amplitude from its average value) improved signal-to-noise ratio can thus be obtained by limiting the range of frequencies received and by increasing the amplitude variation at the transmitter and simultaneously reducing the sensitivity of the receiver to changes in amplitude. Analogously with frequency modulation (in which intelligence is carried by variations of the frequency from its average value), the interfering effect of weak noises may be reduced by limiting the range of amplitudes received, and by increasing the frequency variation at the transmitter while simultaneously reducing the sensitivity of the receiver to change of frequency.

Some forms, such as ignition noise, occur only in bursts which are very short compared with the period of the signal being transmitted (e.g., in telephony, short compared with the period of the highest audio frequency). It is advantageous in this instance even with amplitude modulation to *widen* rather than *narrow* the frequency band of the receiver and to use limiters; this widening, it is true, permits more noise power to pass into the final detector, but it preserves the shortness of the impulse and permits the limiter to chop off the peak, leaving a pulse no greater than that of the signal and so short in duration as not to be harmful.

FADING. In its most general sense "fading" means a reduction of the signal for any cause, including, for example, the slow decrease in long-wave signal strength called the "sunset minimum," which may last an hour or so. Commonly the term refers to the more rapid variations encountered with medium and short waves. Probably the most important cause of fading is the interference of wave components following different paths in space, combined with variations of phase of one or more components with time. The existence of the different components can be due to multiple reflections or to double refraction due to the earth's magnetic field, and perhaps other causes. The changing with time of the relative phases can be due to change in the ionic density along the path or to changes in the magnetic field. Pulse experiments, which are able to resolve a complex signal into many components, show that, with the best resolution possible, the components themselves fade. Fading is therefore a very complex phenomenon.

An important example of fading occurs just outside the service area of a high-power broadcasting station. With increasing distance the sky wave finally becomes appreciable compared with the rapidly attenuating ground wave. Atmospheric vagaries make the relative phase of the components vary at random, causing the signal level to fluctuate. It may be further complicated by the reception of more than one sky-wave component. (See article 24.)

Rapidity of fading of ionospheric waves increases, in general, with frequency. Below 100 kc the phenomenon is scarcely noticed; such changes as do occur commonly require an hour or so. At 1000 kc the period is of the order of 1 minute, and the amplitude range extreme. At 10,000 kc the fading rate is 1 every few seconds. Fading due to interference between two or more components should show this characteristic, for a fade would then occur once for every change of 1 wavelength in the path difference. This relation, of course, can be only qualitative, since long waves and short waves usually travel over very different paths, are frequently not used at the same time of day, and do not employ the same number of paths.

A distorted frequency characteristic is one of the results of transmission over two or more interfering paths. Thus, if the times required for the waves to travel over the two paths differ by τ, and if the radio frequency is slowly varied, consecutive maxima and minima will be found, the frequencies of the maxima differing from one another by integral multiples of $1/\tau$. If $\tau = 0.0005$ second, a typical value, the maxima will be separated by multiples of 2000 cycles per second, and even the a-f characteristic may be seriously affected. There will now be fading provided that a change occurs in the medium. Being a function of radio frequency, such fading is called "selective." If τ were made very small, the frequencies of the maxima would differ so much that within the small band occupied by a telephone channel the response would be independent of frequency. This would be nonselective, or "general," fading. The selectiveness of fading is thus associated with the time difference over the two paths, whereas the rate of fading is related to the rate of change with time of path difference measured in wavelengths. It is common to have more than two components in a received wave. In some cases the wave is extremely complex. [Potter, *Proc. I.R.E.*, Vol. 18, No. 4, 581 (April 1930).]

As shown by Bown, Martin, and Potter [*Proc. I.R.E.*, Vol. 14, No. 1, 57 (February 1926)], this mechanism can produce serious distortion in a speech channel if the instantaneous frequency of the transmitter varies during the audio cycle.

The change in field strength with frequency at a single receiving location has its counterpart in a change with location for a constant frequency. The different paths differ not only in length but also in direction. Most important perhaps is the direction in the vertical plane, as shown in Fig. 7, but the directions in the horizontal plane are also significant. It is these directional differences which cause the difference in signal levels at nearby points, for the two waves give rise to a set of interference fringes. Fading does not therefore occur simultaneously at nearby points, and it is found in the high-frequency range that points separated by 10 wavelengths usually fade in an unrelated manner, whereas in some cases the separation need be no greater than 2 or 3 wavelengths. Advantage of these facts is taken in "diversity systems" of reception, which employ receivers operating on the same frequency from separate receiving antennas at different locations or different polarizations, or on different frequencies from the same antenna. [Beverage and Peterson, p. 531, and Peterson, Beverage, and Moore, *Proc. I.R.E.*, Vol. 19, No. 4, 562 (April 1931).]

The effect of fading is to degrade the performance of a circuit used in communication. This can be due merely to reduction of field during the fades, which leads to an inadequate ratio of the signal strength to the noise. It can be the result of the distorted a-f characteristic which was mentioned above; to the production of distortion products, as for example when the carrier in a double sideband system fades out, leaving the two sidebands to beat with each other; to the existence of fading so rapid that it cannot be compensated by such devices as the automatic volume control, and to other complications. Probably the most usual example of circuit impairment occurs when fading and noise contribute simultaneously. All such effects are the more serious the higher the standard set for the circuit.

SOLAR DISTURBANCES. Radio transmission is one of the terrestrial phenomena which may be correlated with solar activity; another is the variations which occur in the earth's magnetic field, extreme fluctuations having been given the name "magnetic storms." Except for daylight transmission with low frequencies (e.g., 60 kc) the effect of unusual solar activity is an adverse one. By day, low frequencies are then somewhat aided, but at night their fields are considerably reduced. The most marked effect, however, is produced on the high frequencies; in fact, very severe storms may completely eliminate their usefulness over some paths. These effects are illustrated in Fig. 13, which shows daytime field strength on two transatlantic channels, one (full line) on 18,000 kc and the other (heavy dashes) on 60 kc, together with the horizontal component of the earth's magnetic field (light dashes). [Anderson, *Proc. I.R.E.*, Vol. 17, No. 9, 1528 (September 1929).] A striking demonstration that rapidly moving ionic clouds are hurled into the ionosphere during magnetic storms has been given by Wells, Watts, and George of the Carnegie Institution of Washington by a technique of rapidly recording the reflections received over an extended frequency range.

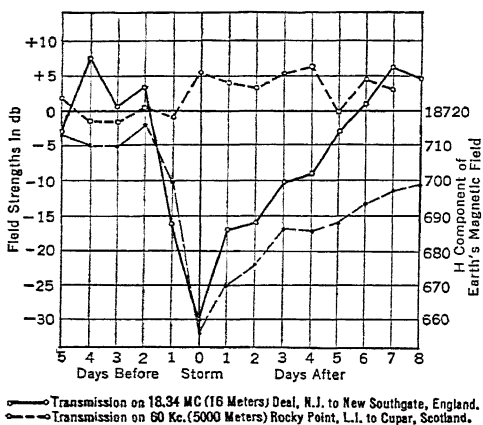


FIG. 13. Typical Effects Accompanying Magnetic Storms

The solar phenomena include sunspots, prominences, and flocculi which may be observed with a telescope or spectrohelioscope. The disturbing areas rotate with the sun once in every 27 days, this being the reason for the ill-defined "period" of 27 days in the terrestrial effects. In order to have a terrestrial effect it seems to be necessary that the disturbed solar area have a certain orientation with respect to the earth. The 27-day period in magnetic and radio effects forms the basis for a method of predicting future disturbances; these predictions, though not entirely reliable, are useful. Another periodicity in solar activity is the secular one, having a cycle of about 11 years. Since the last minimum and maximum occurred respectively in 1944 and 1947, it appears that the next will take place about 1954 and 1958.

Depending on the severity of the "storm," the effect may last one or several days, during which communication on short waves is erratic and difficult and sometimes impossible. At the same time, the aurora may be visible, earth potentials usually rise to relatively high values, and the earth's magnetic field may be seriously disturbed. Short-wave transmission along paths through equatorial regions are scarcely affected, however; it is in the auroral zone that the effects are produced, and, apparently, high-frequency transmission over any path, long or short, which requires reflection from the ionosphere in this zone, is adversely affected. At such times, experiments to determine virtual heights in these zones are impossible owing to the total absence of reflections. [Appleton, Naismith, and Builder, *Nature*, Vol. 132, No. 3331, 340 (Sept. 2, 1933).] The secular period in this effect has already been mentioned. Figure 14 illustrates this variation, which follows the 11-year sunspot cycle. [Austin, *Proc. I.R.E.*, Vol. 20, No. 2, 280 (February 1932).]

Another phenomenon associated with the sun is the radio fade-out, during which all sky waves except those of low frequency are suddenly weakened or obliterated over the earth's sun-lit hemisphere. It has been established that the fade-out is coincident with a bright solar chromospheric eruption, and the absence of the effect at night and its great intensity at the equator indicate ultraviolet light rather than particles as the means by which it is produced. A fade-out may last from a few minutes to a few hours. It is evidently caused by an unusually high electronic density produced below the *E*-region which operates by absorption due to the high collision frequency with neutral molecules at that level. [J. H. Dellinger, *Science*, Vol. 82, No. 2128, 351 (Oct. 11, 1935); L. V. Berkner, *Phys. Rev.*, Vol. 55, No. 6, 536 (March 15, 1939).]

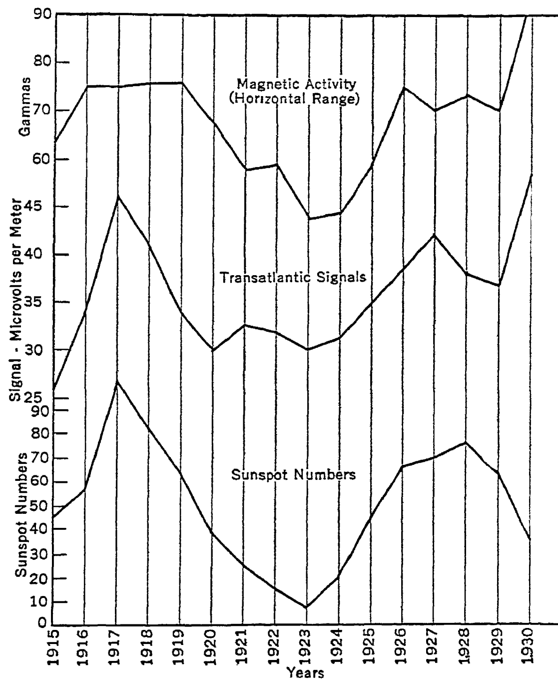


FIG. 14. The "11-year" Period in Sunspot Activity and Its Correlation with Magnetic Activity and Low-frequency Radio Transmission

An abnormality known as sporadic *E* layer reflections is of some importance as the cause of occasional long-distance transmission at frequencies above the usual ionospheric limit, sometimes as high as 60 mc. It occurs in patches rather than uniformly over the *E* layer and is not well understood. [L. V. Berkner and H. W. Wells, "Abnormal Ionization of the *E*-Region," *Ter. Mag. and Atmos. Elec.*, Vol. 42, No. 1, 73 (March 1937).]

24. RANGE OF RADIO STATIONS AND BROADCAST COVERAGE

The distance over which communication can be carried on, or the "range" of a radio transmitting station, depends on so many changeable phenomena and special details that the term is usually significant only as an order of magnitude or as a statistical mean. The range depends on transmission efficiency and noise (e.g., atmospherics), on the types of apparatus used at the receiving station, and on the standards of performance. More meaning attaches to the range of a ground wave, however, because of its steadiness relative to the ionospheric wave. The service range of a broadcast station may be set by noise due to atmospherics, to industrial or domestic electrical equipment, or to unavoidable random noise arising in the receiver itself. But even when noise is negligible it may be limited by fading due to the sky wave's being appreciable compared with the ground wave.

Considering noise as the limiting factor, we may start with representative noise data such as those given in Fig. 15. This figure gives approximate noise values (atmospherics, set noise, etc.) which we assume as typical for medium-frequency broadcast reception in northeastern United States in the summer. For reception the signal field must be greater than the noise by certain values which depend on the grade of service desired. This leads to a required signal field, and the distance at which it is obtained under certain conditions can be obtained by reference to Fig. 3. For example, to find the summer night range of the ground wave of a 100-kw transmitter operating on 1500 kc, we have:

Summer midnight noise on 1500 kc, Fig. 15	0.035 mv/meter
Signal-to-noise ratio assumed	100
Field required with 100 kw radiated	3.5 mv/meter
Field with 1 kw radiated $(3.5/\sqrt{100})$	0.35 mv/meter
Distance giving 0.35 mv/meter (51 db above 1 mv/meter) with 1 kw, $\sigma = 2 \times 10^{-13}$, Fig. 3	53 miles

Figures such as these, which are based on transmission data applying in the case of level terrain, cannot in general apply if there are large obstacles or other irregularities in the path. A striking example of this was described by Bown and Gillette, who found that sections of New York City in which there are large numbers of unusually tall buildings cast "shadows" for several miles [*Proc. I.R.E.*, Vol. 12, No. 4, 395 (August 1924)]. Dense areas of small buildings also reduce the range. There are other factors than noise, such as fading and interference from other stations, that complicate this problem and require considerable experience to assess. Field strength surveys of the actual territory to be covered are frequently required. [Standards of Good Engineering Practice, F.C.C., loc. cit.; R. F. Guy, loc. cit.; "An Analysis of Continuous Records of Field Intensity at Broadcast Frequencies," Norton, Kirby, and Lester, *Proc. I.R.E.*, Vol. 23, No. 10, 1183 (October 1935); "On the Use of Field Intensity Measurements for the Determination of Broadcast Station Coverage," Jansky and Bailey, *Proc. I.R.E.*, Vol. 20, No. 1, 62 (January 1932).]

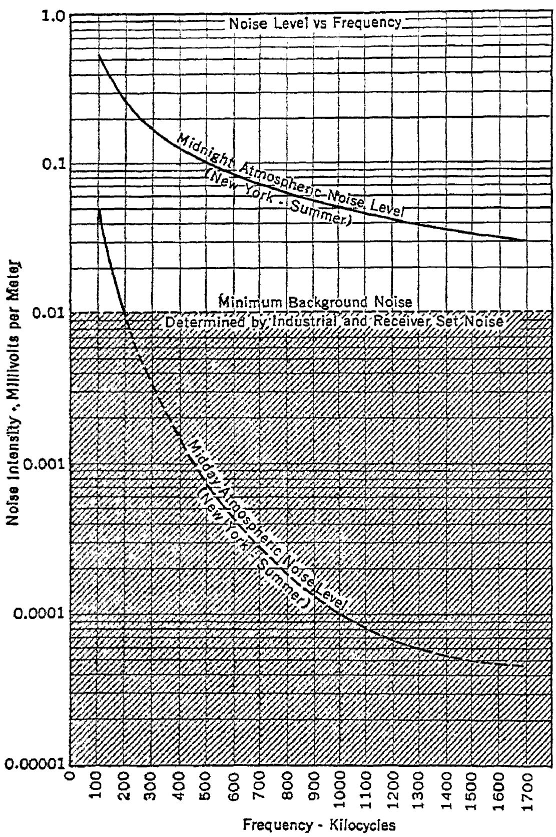


Fig. 15. Typical Noise Data, Broadcast Reception

Another limitation in the service area of a broadcast station in the evening is the existence of fading beyond a certain distance. This limiting distance depends on frequency and earth conductivity, and in general it is greater the lower the frequency and the higher the conductivity. Fading first becomes serious when the sky wave becomes appreciable compared with the ground wave. Figure 16 illustrates, as a function of frequency, how the distance range of broadcast stations depends on electrical noise and fading, an earth conductivity of 10^{-13} being assumed. [Report of Committee on Radio Propagation Data, *Proc. I.R.E.*, Vol. 21, No. 10, 1430 (October 1933).]

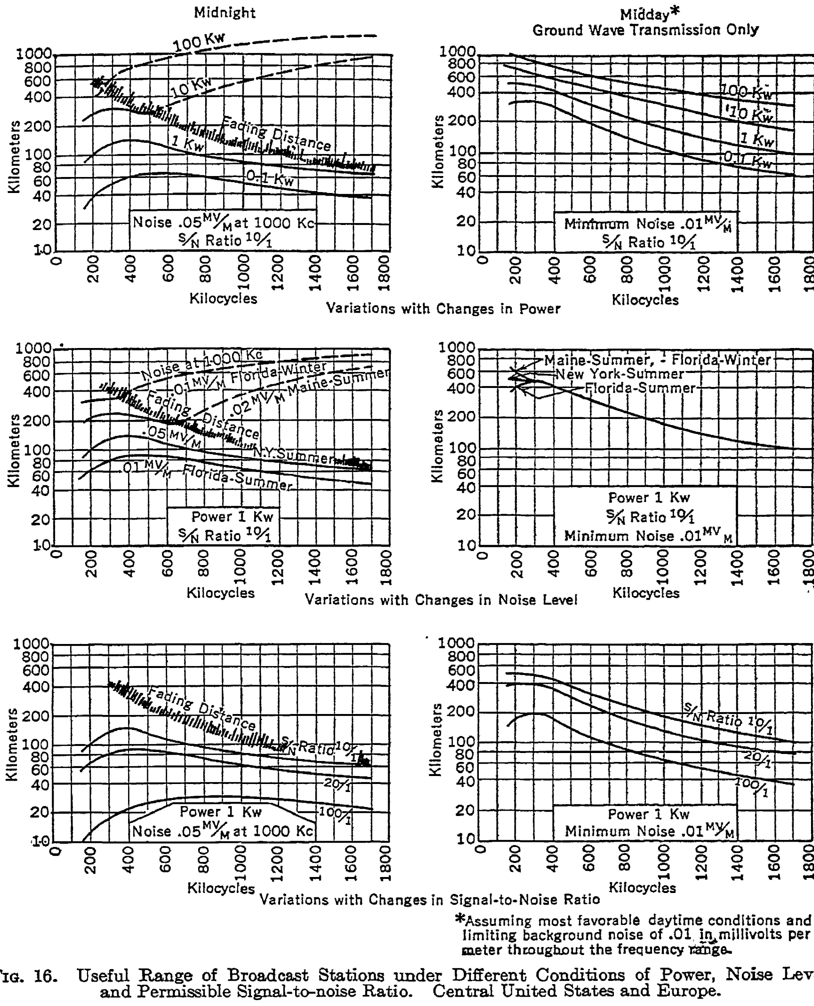


FIG. 16. Useful Range of Broadcast Stations under Different Conditions of Power, Noise Level, and Permissible Signal-to-noise Ratio. Central United States and Europe.

MECHANICAL FEATURES OF TRANSMISSION LINES

By John D. Taylor

25. TRANSMISSION-LINE CONSTRUCTION

POLE LINES are employed in aerial communication construction to support open wires and cables used in toll and exchange plant. The supporting structures are generally of wood but, for special requirements or where pole timber is not obtainable, may be of steel or other materials.

The use of wood poles generally throughout the United States and other countries is due principally to (1) availability, (2) economical type of construction, (3) ease of handling and maintaining, and (4) relatively long life.

The design of wood-pole lines is based primarily on (1) type of communication plant (toll or exchange) to be supported, (2) load to be carried, and (3) location and exposure to weather. The poles must be of sufficient strength (allowing for ground decay and economical life) and height (allowing for ultimate loads and required clearances) to meet the requirements for this type of construction in the most economical manner.

In exchange plant, initial pole-line routes largely determine the routes of distribution for succeeding types of construction, while for toll plant the large initial costs involved in building a toll pole line usually require its maintenance on the selected route, at least for its economical life or until other considerations such as right-of-way, growth, new developments or deterioration necessitate its removal or replacement. In planning new pole lines or rearrangements, it is important to so advise other pole-using companies in the area involved, in order that the plans of all the companies may be in coordination at all times.

The selection of the pole route usually entails advance surveys, acquiring the necessary right-of-way, and other considerations, which will provide the required pole line economically. Due regard must be given to the future adaptability and relation of the pole line with respect to the telephone system, serving the area, as a whole. Toll pole lines usually take the most direct, practicable route between terminating points, avoiding small towns, trees, and hazardous conditions as far as possible. Toll points, off the main toll route, are reached by branch (spur) lines. Along highways, one side should be occupied throughout, as consistently as conditions permit, avoiding unnecessary road crossings and leaving the other side of the highway for other wire-using companies.

Public right-of-way is less expensive initially, but private right-of-way for a particular section of pole line may result in lower annual charges and ultimate costs and most certainly will add to the permanence and safety of the line.

Joint use is generally desirable and economical in urban areas, rather than using separate pole lines for power and telephone facility distribution. The power circuit voltages in cities are usually low (not over 5000 volts), and the telephone equipment and subscribers are adequately protected in case of contact between the power and telephone circuits.

Joint use with rural or toll open-wire circuits is not, as a rule, desirable, because of the generally higher-power circuit voltages and of possible hazards to life and property from contacts. Consideration may be given, however, to such joint use in any particular case, and future developments may indicate its desirability for rural construction.

The selection of poles required for any particular pole line is based mainly on (1) the number of telephone or other aerial wires and cables to provide facilities over the expected service life of the poles, (2) the importance of these facilities, (3) the pole strength required to carry the initial and ultimate wire and cable loads under the weather conditions expected in the locality involved, and (4) governmental regulations.

Some companies have established classifications for open-wire and cable pole lines, in accordance with the service value of the line (relative importance and number of messages carried by the circuits on the line). Figure 1 shows the classifications used by the Bell System and the relative strengths assigned based on the System's experience. Lower percentages of ultimate stress are required for railroad, power, or similar crossings.

Classification of Line	Type of Service	Message-miles per day	Relative Strength Levels	Maximum Percentage of Ultimate Fiber Stress under Transverse Loading	
				For New Poles	At Replacement
A	Toll open-wire and cable of high service value..........	Over 30,000	100	45	67
B	Toll open-wire (average service value) and toll cable (not class A)	5,000 to 30,000	67	67	100
C	Exchange open-wire (over 10 wires), all exchange cable...	....	50	89	133
	Toll open-wire of low service value	Less than 5,000	50	89	133
R	Exchange open-wire (not over 10 wires)................	....	33	133	200
J	Joint power and telephone lines....................	....	88	50	75

FIG. 1. Classification and Strength Requirements of Pole Lines, Bell System Practices

In connection with the preparation of the National Electrical Safety Code (N.E.S. Code), National Bureau of Standards Handbook H32, studies were made to determine the frequency, severity, and effects of ice and wind storms throughout the country. On the basis of these studies, three general loading areas, *heavy, medium, and light,* were established for the United States, as shown in Fig. 2. For the same classification of line and other structural conditions, a heavier class of pole is required (particularly for average or greater loads) in the heavy than in the medium or light loading areas, and likewise in the medium as compared to the light loading area.

Basic conductor loadings have been assigned for the three loading areas, in order to derive pole loadings, considered appropriate for these areas, and with the various pole-line classifications and other data, to arrive at the class of pole required for any given line. Figure 3 shows these assumed loadings and associated constants, which latter, when added to the resultant of the vertical and horizontal loads, will result in effective conductor loadings, substantially the same for the Fifth Edition as for the Fourth Edition of the N.E.S. Code. Thus, it has been possible to avoid lowering past overall effective standards for pole-line strength requirements, and at the same time to reduce, in the Fifth Edition, the transverse loadings on the pole line to permit the use of allowable stress values more nearly representative of general engineering practices.

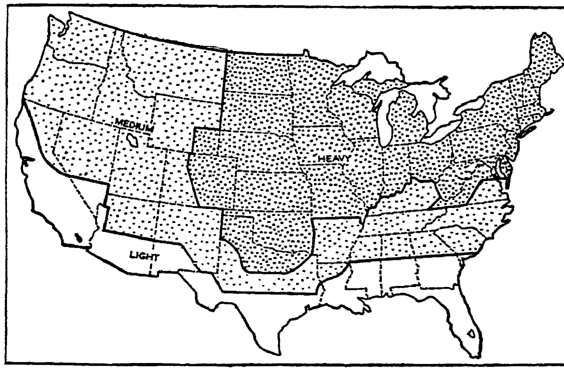


FIG. 2. Storm Loading Map of the United States (from N.E.S. Code, Fifth Edition, *Natl. Bu. Stds. Handbook H32*)

Assumed Vertical and Transverse Loadings

Storm Loading Area	Symbol	Radial Thickness * of Ice Coating on Conductors and Messengers, in.	Horizontal Wind Pressure at Right Angles to the Line, lb/sq ft of projected area
Heavy.........	H	0.50	4
Medium.......	M	0.25	4
Light..........	L	None	9

* *Note:* In computing transverse loading on poles and towers, ice coating on these structures is ignored.

Constants for Various Types of Conductors to Be Added to the Resultants of the Loadings Shown in above Table, pounds per foot

Storm Loading Area	Symbol	Temperature, deg F	Bare Copper, Steel, Copper Alloy, Copper-covered Steel, and Combinations Thereof	Bare Aluminum with or without Steel Reinforcement	Weather Proof and Similar Covered Conductors (All Materials)	Cable with Messenger
Heavy..........	H	0	0.29	0.31	0.31	0.6
Medium........	M	+15	.19	.22	.22	.4
Light..........	L	+30	.05	.05	.05	.2

Note: For telephone wires it is usually assumed that

$$P = 0.003V^2$$

where P = horizontal wind pressure in pounds per square foot of projected area.
V = actual wind velocity in miles per hour.

FIG. 3. Assumed Vertical and Transverse Loadings and Associated Constants (from N.E.S. Code, Fifth Edition, *National Bureau of Standards Handbook H32*)

Since pole lines carry various types of communication facilities and frequently (on joint lines) power conductors and equipment as well, it is necessary, in determining pole-line loads, to equate the various attachments to a common basis. Figure 4 gives wire equivalent data for the three loading areas in terms of effective 104 (mil diameter) telephone wires and of effective No. 4 covered power wires. In heavy and medium loading areas only,

Attachment	Telephone Wire Base (Effective 104 Tel. Wires)			Power Wire Base (Effective No. 4 Covered Power Wires)		
	Storm Loading Area			Storm Loading Area		
	Heavy	Medium	Light	Heavy	Medium	Light
Communication plant						
Bare open wire, 109, 104, or smaller, per wire	1	1	1	0.8	0.7	0.3
Bare open wire, 128, 134, or larger, per wire	1	1	1.3	0.8	0.7	0.4
Covered paired wire, per pair, or covered single wire, per wire	1	1	2.5	0.8	0.7	0.7
Strand, all sizes	2	2	5	2	2	2
Cable and 6000 lb (5/16 in.) strand	4	5	15	3	3	4
Cable and 10,000 lb (3/8 in.) strand	5	6	22	4	4	6
Cable and 16,000 lb (7/16 in.) strand	6	7	30	5	5	8
Cable and 25,000 lb (1/2 in.) strand	7	8	35	6	6	10
Cable terminal, "B" or "BB" type, 202 pair and less or "BD" type, all sizes	1	1	4	1	1	1
Cable terminal, "B" or "BB" type, more than 202 pair	1	1	6	1	1	2
Cable loading pot	1	1	4	1	1	1
Service drops, per unbalanced drop	1	1	2	1	0.5	0.5
Clothes-lines on Class C line poles, per unbalanced clothes-line	8	14	50			
Clothes-lines on Class J line poles, per unbalanced clothes-line	4	7	25	3	5	8
Power plant						
Covered wire, No. 8 AWG (approx. 0.26 in. o.d.) or smaller, per wire	1.1	1.3	2.5	0.9	0.9	0.7
Covered wire, No. 6 AWG (approx. 0.32 in. o.d.) per wire	1.2	1.4	3.1	1.0	1.0	0.9
Covered wire, No. 4 AWG (approx. 0.38 in. o.d.) per wire	1.3	1.5	3.7	1	1	1
Covered wire, No. 0000, AWG (approx. 0.65 in. o.d.), per wire	1.5	1.9	6.3	1.2	1.3	1.7
Covered wire, 500,000 circ mils (approx. 1.11 in. o.d.), per wire	2	3	11	2	2	3
Covered wire, 1,000,000 circ mils (approx. 1.53 in. o.d.), per wire	3	4	15	3	3	4
Covered wire, 2,000,000 circ mils (approx. 2.15 in. o.d.) or larger, per wire	3	5	20	3	4	6
Power cable on strand (approx. diam. of cable 2.56 in. or less)	6	7	30	5	5	8
Suspension wire extending transversely between two pole lines and supporting trolley contact wires, per line: One contact wire	6	10	25	5	7	7
Two contact wires	12	20	50	10	14	14
Four contact wires	18	30	75	15	21	21
Bracket and one trolley contact wire on one side of pole line	2	2	8	2	2	2
Brackets and two trolley contact wires, one on each side of pole line	3	3	9	3	2	3
Bracket and two trolley contact wires, over tracks on same side of pole line	5	6	19	4	4	5
Transformers, 37 1/2 kva or less	1	1	6	1	1	2
Transformers, over 37 1/2 kva	1	2	9	1	2	3
Transverse clearance attachment for service drop above telephone attachments, per wire	1	2	4	1	2	1
Service drops, per unbalanced drop wire	1	1	3	1	1	1
Street lamp supported by mast-arm (not bracket)	1	1	3	1	1	1

Fig. 4. Table of Wire Equivalents for Pole-line Loading Calculations (from N.E.S. Code, Fifth Edition, *National Bureau of Standards Handbook H32*)

when the actual number of wires on crossarms is more than 10, shielding reduces the effective number of wires to 67 per cent of the actual number. In light loading areas, loading is assumed on actual wires.

The selection of new poles for a given pole line requires that the following factors be known or assumed:

(a) Classification of pole line (determines the relative importance of the line).

(b) Loading area (determines the basic loadings).

(c) Fiber strength of timber to be used (see Table 1).

(d) Equivalent (effective) telephone wire or power wire load (see Fig. 4).

(e) Average pole spacing (depends on type and size of wire, usage, loading area, and class of line for open wire).

(f) Length of poles (determined by load and clearance requirements).

Table 1

Type of Timber	Fiber Strength, lb/sq in.
Northern white cedar	3600
Western red cedar	5600
Creosoted southern pine and Douglas fir	7400
Chestnut	6000
Lodge pole pine	6600
Juniper	4600
Cypress	5000

If tables are not available for determining directly the class of poles required under the above known or assumed conditions, the following formulas may be used for the purpose.

For *transverse loading* (pole acting as a cantilever beam)

$$C = \sqrt[3]{\frac{M_w + M_p}{0.000264fU}}$$

where C = minimum required ground line circumference of new pole in inches.

M_w = resistant moment on pole at ground line, due to wind pressure on wires, in ft-lb.
$= PDLNS.$

P = wind pressure, in lb/sq ft.

D = diameter of each wire, in ft, including ice coating, if any.

L = distance, in ft, between center of the load and the ground line section.

N = effective number of wires on the pole.

S = $^1/_2$ sum of the two adjacent spans, in ft.

M_p = resistant moment on pole at ground line, due to wind pressure on pole, in ft-lb.

$$= P\frac{h^2}{226.2}(2C_t + C_g)$$

where h = height of pole above ground, in ft.

C_t = circumference of pole at top, in in.

C_g = circumference of pole at ground line, in in.

U = maximum percentage of ultimate fiber stress expressed as a decimal (determined from the line classification, Fig. 1, or as required).

f = maximum allowable fiber stress of the pole timber, in lb/sq in.

For *vertical loading* (usually considered only for anchor guyed poles or stubs, where the vertical component of the stress in the guy may be large), the pole is considered to be a long column, and by Euler's formula for such a column

$$P_v = \mu E \frac{I}{l^2}$$

where P_v = vertical load on the pole, in lb.

$\mu = \pi^2$ for average conditions of a guyed pole or stub.

E = modulus of elasticity of the pole timber.

I = moment of inertia of the critical section (see note).

l = length of the column, in in., from the point of anchor guy attachment to the butt of the pole (for poles set in solid bases, such as rock or concrete, the length is taken from the point of anchor guy attachment to the ground line and $\mu = 2\pi^2$).

Note: The critical section in flexure for the average pole (of conical shape) is assumed to be at a distance of $^1/_3\, l$ below the point of anchor guy attachment. The circumference of this section should be computed from (1) its distance above a point, which is 6 ft from the butt of the pole, (2) the specification circumference 6 ft from the butt, and (3) the average circumferential taper for the timber under consideration.

In general, the size of line pole, selected for the ultimate load, is determined by the transverse stress, to which the pole may be subjected. There may be cases, however, where the vertical load, particularly where large transformers are mounted near the top of the pole, becomes a substantial factor in determining the pole strength required.

In such cases, both transverse and vertical stresses and their resultant stress may need to be determined, although the vertical and combined stresses can only be calculated approximately.

A wood pole is considered as a long tapering column, which, under a critical concentric vertical load, fails by buckling. The maximum load that the pole can safely carry depends on the degree of freedom of the pole ends, the distribution and arrangement of the vertical loading on the pole, the support given the pole by any wire or cable attachments, and other variable factors.

For wood, the maximum compressive strength is considerably less than the modulus of rupture and the maximum allowable stress for the combined compression and bending stress is intermediate between these two values.

Methods have been developed for determining approximately the combined stress on a pole at the ground line, due to vertical (axial) and transverse loads, under certain assumptions, but, because of their complexity, these methods are not discussed in this handbook.

Where it is desired to employ such methods, reference may be made to *Provisional Report No.* 24 (*Technical Report* 2G-2), A Study of Pole Strength in Jointly Used Poles, of the Joint Subcommittee on Development and Research of the E.E.I. and B.T.S. of August 22, 1938.

The minimum required ground-line circumference of the pole C having been determined, the class of pole actually selected from the ASA Specification Tables should have a ground-line circumference at least equal to the minimum required ground-line circumference C plus an amount of wood which, based on average decay rates for the timber and location involved, will provide the desired service life in accordance with the formula

$$C_s = C + y(L - T)$$

where C_s = minimum ground-line circumference of the new pole, in in., to provide desired service life.

C = minimum required ground-line circumference of new pole, determined as above.

y = average rate of decay of untreated timber in equivalent inches of circumference per year. (For the cedars and chestnut, this decay is about 0.45 in. for sapwood and 0.3 in. for heartwood per year. For southern pine the physical life is about equal to the effective period of treatment.)

L = desired physical life in years.

T = expected life of preservative treatment in years (probably roughly 20 years for initial treatment of butt-treated poles).

The ASA (American Standards Association) Specification Tables of Pole Dimensions, referred to above, are readily available and are based primarily on:

1. Fiber strengths of various timbers used.
2. Ten classes of poles (1 to 10) with minimum circumferences for 6 ft from the butt specified for classes 1 to 7, and minimum circumferences at the top of the pole specified for all classes.
3. Breaking loads 2 ft from the top for the first seven classes in approximate geometric progression as follows:

Class	Breaking Load, in lb
1	4500
2	3700
3	3000
4	2400
5	1900
6	1500
7	1200

4. All new poles of the same class and length (classes 1 to 7 only) to have about equal resistant moments at the ground line.
5. All new poles of the same class (classes 1 to 7 only) to be of such size as to have about the same breaking load, with the load applied 2 ft from the pole top and assuming that the break would occur at the ground line.

Treatment of poles to prolong their physical life is standard practice. The preservatives used may be creosote, greensalt, or other chemicals toxic to wood-destroying fungi and wood-boring insects. The creosote treatment may be applied the full length of the pole, as in southern pine, under a pressure and vacuum with a net retention normally of 6 to 8 lb of creosote per cubic foot of wood, and a penetration of not less than 2 1/2 in. or 85 per cent of the sapwood. Greensalt is also applied to poles by the pressure method, and it has some advantages over the creosote treatment. Butt-treated poles, such as the cedars, are usually processed in open tanks of hot and then cold creosote, the ground-line section being incised to assist penetration, which averages about 0.4 in. or more.

The spacing of poles depends largely upon the type of the wire or cable load, location, transposition scheme, and exposure to storms. For important backbone toll routes the spacing is generally about 130 ft for open wire and from 150 to 300 ft for toll cable.

The guying of pole lines is necessary at points of above average stress, such as at corners having substantial pulls, dead-end poles, and poles carrying unusual loads. Guying is also applied on toll lines at periodic intervals along the line to assist the pole-line structure in withstanding storms. The ratio of lead to height of guys should be about 1.0 to 1.25, the guy stress then being about 1.4 to 1.28 times the horizontal stress.

Cross-arms are designed to carry from 2 to 10 or 12 wires, as may be required for any particular case, and may be fitted with locust or steel pins, usually spaced from 6 to 12 in. for telephone and 10 and 11 1/4 in. for telegraph wires (16 to 30 in. for pole pairs). The cross-arms are usually spaced 24 in. apart vertically on the pole and vary in number per pole from 1 to 6 or more.

OPEN WIRE, supported on insulator-equipped cross-arms or brackets, which are in turn mounted on poles or fixtures, has been employed since the invention of the telegraph and telephone to connect individual instruments to wire centers or offices, and one office to another. However, wire conductors enclosed in lead sheaths (cables) have practically superseded open wire in built-up communities and cities and in large measure, for telephone communication, have replaced or supplemented open-wire lines between the principal cities.

The types of open wire employed for telephone communication circuits, as discussed in Section 17, consist principally of 104, 128, and 165 hard-drawn copper and some 104 and 128 copper steel (40 per cent cond.) for toll circuits, 080 copper steel (40 per cent cond.), and 080 and 109 high-tensile steel for exchange circuits in outlying areas (all given in mil diameter). High-tensile steel wire, because of the pole economies realized, its greater strength, and its equally good service performance, is being employed for new construction generally in place of the various grades of mild steel and iron wire.

The types of open wire employed for telegraph communication circuits consist mainly of 114 hard-drawn copper, 162 copper steel (40 per cent cond.), and some 165 iron for important facilities.

Drop wires of various types are used for both services.

Table 2 shows some of the important physical properties and the electrical resistance of wire, classed as open wire, for both telephone and telegraph circuits. Other electrical characteristics of various types of wire are discussed in Section 17.

Table 2

Type of Wire	Wire Number and Gage	Nominal Wire Diameter, in mils	Average Weight per Wire, in lb/mi	Minimum Breaking Strength per Wire, in lb	Average Resistance per Loop Mile, in ohms at 68° F
Telephone					
Hard-drawn copper..........	14-NBS	80	102	330	17.50
Hard-drawn copper..........	12-NBS	104	173	550	10.15
Hard-drawn copper..........	10-NBS	128	262	819	6.74
Hard-drawn copper..........	8-BWG	165	435	1325	4.11
Copper steel (40%)..........	14-NBS	80	96	770	42.8
Copper steel (40%)..........	12-NBS	104	159	1177	25.0
Copper steel (40%)..........	10-NBS	128	240	1647	16.7
High-tensile steel (HTL-85)					
(0.8 oz zinc coating).....	14-BWG	83	99	460	117.2
(HTL-135)...............	14-BWG	83	99	703	130.0
(HTL-85)...............	12-BWG	109	170	793	68.2
(HTL-135)...............	12-BWG	109	170	1213	76.5
(HTL-85)...............	10-BWG	134	258	1199	45.0
Bronze TP drop............	18-AWG	40	159	340	259.0
Bronze TR drop............	18-AWG	40	232	340	259.0
Bronze NP drop............	18-AWG	40	227	340	259.0
Hard-drawn copper HC drop.	14-AWG	64	316	380	26.4
Telegraph					
Hard-drawn copper..........	9-AWG	114	208	644	4.3
Copper steel (40%)..........	6-AWG	162	384	2430	5.3
Iron......................	8-BWG	165	378	1090	13.3
Copper (tw. pr.)............	16-AWG	51	208 (pr.)	500 (pr.)	104.0 (pr.)
Steel drop (sgl.)............	16-AWG	51	104 (sgl.)	250 (sgl.)	52.0 (sgl.)

CABLE, consisting of insulated, annealed copper conductors, enclosed in a cylindrical lead sheath, is employed in both toll and exchange telephone and in telegraph plant, where the number of circuits required along a given route, plant economies, or interfering conditions preclude the use of open wire. Cable is used generally in toll plant in urban areas and, supplementing open wire, between principal traffic centers throughout the country, either as toll entrance facilities for open-wire lines or as toll cable facilities directly connecting large switching centers.

Telephone and telegraph cable is manufactured in various sizes and gages, of which representative types and associated data are given in Table 3. For both types of service,

Table 3. Representative Cables—Mechanical Characteristics

Type	(3) No. of Pairs	(1) (2) Gage of Conductor	Conductor Insulation	Sheath Thickness, in.	Outside Diameter, in.	Weight per Foot, lb	Type of Core
Exchange.........	6	19	Paper tape	0.063	0.42	0.41	Layer
Exchange.........	455	19	Paper tape	0.115	2.61	8.48	Layer
Exchange.........	11	22	Wood pulp	0.063	0.42	0.40	Layer
Exchange.........	909	22	Wood pulp	0.115	2.61	8.46	Multiple-unit
Exchange.........	11	24	Wood pulp	0.061	0.36	0.31	Layer
Exchange.........	1515	24	Wood pulp	0.115	2.61	8.64	Multiple-unit
Exchange.........	11	26	Wood pulp	0.060	0.32	0.27	Layer
Exchange.........	2121	26	Wood pulp	0.115	2.61	8.15	Multiple-unit
Toll entrance......	(4) 12 quads	19	Paper tape	0.082	0.85	0.79	Layer
Toll entrance......	3 quads 27 quads 60 quads	13 16 19	Paper tape Paper tape Paper tape	0.118	2.35	6.41	Layer
Toll (full size).....	154 quads	19	Paper tape	0.123	2.59	7.60	Layer

Notes: (1) All conductors are annealed copper.
(2) Some 28 gage exchange cable was made during World War II to conserve copper in sizes 11 to 303 pairs.
(3) The smallest and largest sizes of exchange cable are shown, as used in the Bell System.
(4) Toll entrance and toll cable may have optional groups of non-quadded exchange pairs. The cable data given assume no shielding or sheath protection or exchange conductors.
(5) Insulation resistance required to exceed 500 megohms per mile.
(6) Some 13 gage is used in both telephone and telegraph service.

paper or wood-pulp insulated conductors are grouped together to form a core, either in layers or 51 and 101 pair units, various colors being used in the insulation and binding strings to permit readily distinguishing between different layers, units, quads, or pairs, for installation or maintenance purposes. Figure 5 shows the method of core construction for both layer and multiple-unit type cores of telephone exchange cable (24 gage-type DSM).

Exchange conductors are generally associated in pairs (2 conductors twisted together), although some exchange trunk cable is quadded (2 pairs twisted together as a 4-conductor unit group, called a quad). Toll conductors are usually quadded for phantom circuit operation, although toll cables frequently contain some non-quadded pairs for program or other toll services. In addition, complements of exchange pairs are frequently included in the toll cable. The lead sheath enclosing the paper-wrapped core of insulated conductors normally contains about 1 per cent of antimony to strengthen the sheath, the thickness of which varies with core diameter and type of cable from about 0.06 to 0.125 in.

Cross-talk must be carefully considered in cable design, particularly for toll cables. In non-quadded cables the conductors are twisted together in pairs. The twists in adjacent pairs within a layer and in adjacent layers are of different lengths to provide the necessary capacitance balance between pairs. In quadded cables the conductors are twisted together in pairs and the pairs are twisted together with different lengths of twist to form quads of as many as 9 types, each of which has a different length of twist. Exchange pairs (non-quadded) are usually random spliced within their color groups (spliced without testing to determine the pair numbers), so that any given pair in the cable is adjacent to any other pair in as few cable sections as practicable. Toll pairs, because of their greater importance, wider range of operating energy levels and frequency assignments, are carefully spliced where sections of cable join. The splicing is carried out according to a plan which provides for limiting the cross-talk between circuits.

Segregation, to reduce couplings at carrier frequencies between oppositely transmitting pairs or quads used for cable carrier, is usually accomplished by assigning oppositely bound groups to separate cables. Segregation of pairs used for open-wire carrier is obtained by using alternate layers or by metallic layer or unit quad shields in the same cable, as may be required to prevent excessive cross-induction between them.

Sheath protection is provided in various degrees, as may be required, by using: (1) layers of Sisalkraft paper and jute covering, where soil corrosion may occur; (2) layers of thermoplastic compound covered by longitudinal copper tape, flooded with asphalt compounds, for a lightning shield; (3) layers of Sisalkraft paper, two steel tapes, and jute, for gopher-infested areas; (4) layers of Sisalkraft paper and a layer of rubber or asphalt-back fabric tape for corrosion protection in conduit; (5) a thermoplastic compound layer and outer covering of impregnated fabric tape for corrosion protection when buried near pipe lines; (6) two helical wrappings of tape armor and a cushion of jute to protect against ring cutting, stone bruises, or to provide shielding for low-frequency induction for aerial cables; and (7) single and double armor wires with a jute cushion for submarine cables. Other special protection may be provided as required.

Spiral-four disk-insulated cable is used, under certain conditions, to provide entrance facilities for open-wire carrier systems operating at a maximum frequency of about 140 kc. The individual units contain 4 copper conductors of 16 gage, insulated from each other by composition disks having uniformly spaced peripheral notches which hold the wires spaced at the corners of a square, each oppositely positioned pair of wires forming a pair of the quad. Shielding is provided over a spiral-4 unit consisting of a copper tape and two steel tapes. A toll entrance cable may contain up to 7 such units, each with its own metal sheath, or a combination of units and standard paper-insulated quads and pairs which provide other communication facilities.

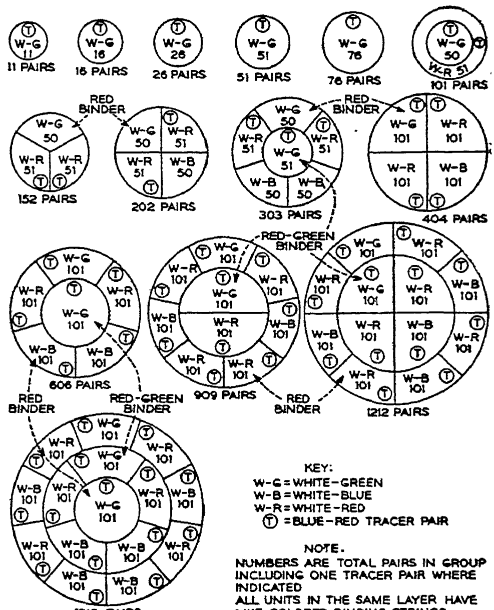


FIG. 5. Diagrams Showing Typical Layer and Multiple-unit Type Exchange Cable Cores (Courtesy Bell System)

Coaxial cable, of latest design, for use with the carrier and other high-frequency systems, consists of up to 8 units. Each unit is composed of a 100.4-mil conductor (inner conductor) positioned in the center of a single 12-mil copper tape (outer conductor) with longitudinal seam to form a 0.375-in. (inner diameter) tube, over which are lightly wrapped two 6-mil steel tapes. The outer conductor is supported by polyethylene disks, 0.085 in. thick, spaced about 1 in. apart along the inner conductor. The outer and inner conductors form a metallic circuit. With the 8 units, other standard paper-insulated quads of conductors (up to about 78 quads of 19 gage or equivalent) may be included in the same overall lead sheath. Further details regarding this cable and its characteristics are given in Section 17.

Both exchange and toll cables are *placed aerially and underground*, depending upon the location and other factors affecting any given cable.

When cables are placed aerially on poles or towers, the required strength of the supporting structures must be carefully determined on the basis of loading area, exposure to weather conditions, load carried, importance of the cable to service, and other factors such as apply in open-wire construction. When cables are placed underground, the main factors to consider are the type of underground housing, if any, to employ, as well as the possibilities of damage from corrosion or other external sources, and the proper routing. Underground cables may be placed in vitrified clay conduit, fiber, or other types of ducts or buried directly in a trench in the ground or plowed into the ground. In any event, suitable protection from underground hazards must be provided.

Loading coils are provided in cable circuits (aerial or underground) wherever required for transmission reasons (see Section 17).

Loading coils are generally assembled in groups in steel cases for either aerial or underground installations, and some designs are suitable for office relay rack mounting. Lead sleeve cases are used for loading small complements of toll and exchange conductors. Also single coils (in a small metal case) are designed to be connected and enclosed in cable splices.

UNDERGROUND STRUCTURES, consisting of conduit, manholes, vaults, and other construction, are designed to provide suitable housings for underground cable plant. These structures are a necessity in cities and to some extent in the smaller communities (assuming that aerial cable plant is not practicable or economical in a given case), since the underground cable plant must be readily accessible to permit additions, changes, removals, and repairs. Some underground exchange cable has been buried (placed directly in the ground) in built-up locations, but this is not the general practice. Toll cables, suitably protected, are frequently plowed into the ground or are buried in trenches. This type of construction has been employed for long distances over transcontinental and other important toll cable routes.

Underground conduit may be of vitrified clay, fiber composition, creosoted wood, or iron pipes. Clay conduit is most generally employed for main routes because of its relatively low cost and satisfactory performance when buried and properly protected. It is manufactured in single or 2-, 3-, 4-, 6-, 8-, and 9-duct multiple units, and is usually laid in one or more units in a compact arrangement. The conduit is placed in a trench, which may vary in depth to avoid other underground structures or hazards (sharp changes in direction are avoided), on a solid base with or without concrete. It may be encased completely in concrete or only at the top or bottom, or plank may be used, in the judgment of the engineer as to the type of construction needed.

Conduit may be placed along a street under paving or the sidewalk, or between the curb and sidewalk, or in parking in the center of the street, as conditions permit. Costs and a satisfactory permanent location, least likely to be disturbed, are primary considerations in placing conduit over a selected route. Subsidiary ducts of wood and of vitrified clay, sewer or iron pipe are commonly used from manholes to underground cable poles or buildings.

Manholes are required at junction points of conduit runs and other locations where it is desirable that the underground cable plant be made accessible, and to provide for practicable cable section lengths. Manholes are preferably located at one side or the other of street intersections to avoid interfering with traffic, when working in them. Manholes vary in size from the small service boxes, placed mainly for the purpose of pulling cable, to the standard 2-, 3-, and 4-way and center rack types, having dimensions ranging from 3 ft 6 in. width by 6 ft length by 5 ft 6 in. headroom to 8 ft width by 9 ft length by about 6 ft or more headroom. In addition, special type manholes and cable vaults of various sizes are required in many cases to accommodate concentrated conduit entrances, as at central offices and loading coil installations. Concrete construction is usually employed, with or without reinforcement. Manhole frames and covers of cast iron, placed in the manhole roof to provide a suitable entrance of 27-in. diameter or more to the manhole, are designed to support surface traffic safely. Drainage may or may not be provided in manholes, as required.

The underground cables are supported on galvanized-metal racks, mounted vertically along the sides or through the center of the manhole, and are separated by a few inches to permit splices to be made and opened as required.

26. ELECTRICAL PROTECTION OF TRANSMISSION LINES

Lightning and power circuits are two unlike sources of electrical power, which, under certain conditions of exposure or contact encountered in practice, occasionally cause damage to communication plant.

Communication lines and equipment are necessarily grouped closely together because of space limitations and other economic considerations. These facilities will carry their normal operating voltages and currents with ample margins of safety, but the insulation employed on cable conductors, between conductors and the cable sheath, and in various equipments is not sufficient, in general, to withstand lightning or power circuit potentials without protection.

EXCHANGE CABLE PROTECTION against lightning for any given exchange cable plant usually requires a study of the plant layout and exposure conditions involved. The aerial cable sheath is usually grounded through the underground cable plant or at the central office and at various other locations at irregular intervals, such as at underground dips or private cable entrances.

The working cable conductors are grounded by operation of the station protection and through the loops connecting to them. Cable conductors may serve subscriber stations at one or more intermediate points along the conductors but lie idle in other sections of the cable beyond these points. Other conductors may lie idle throughout their length, and groups of conductors will terminate at various points where cable sizes change. All these conditions have a direct bearing on potentials in exchange cables due to lightning.

Lightning currents may, in general, be impressed on exchange cables by (1) inductive coupling between the cables (or between conductors connected to them) and the lightning path, (2) direct stroke, or (3) metallic connection with entering open-wire or drop loops, carrying lightning potentials, as a result of a direct stroke or other causes. Lightning currents from metallic connections (by conduction) are usually the principal consideration in the design of exchange cable protection.

Lightning current may appear on the open wire, connected to the cable, from such sources as direct strokes to the line, conduction from service drops or guys, or from arcs from power-system ground wires. Figure 6 shows the results of crest current measurements at exchange cable-open wire junctions.

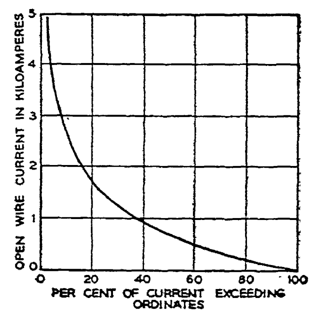


FIG. 6. Crest Current Distribution for Lightning Currents at Exchange Cable—Open Wire Junctions (Courtesy Bell System)

Although lightning currents reach cables over both connected drop and open-wire loops, it appears that the currents on the drop loops will probably be greater than those on the open wire. However, cable plant with drop loops is generally located in built-up areas having low-resistance public water systems and various types of shielding, whereas open-wire plant extends mostly into rural areas where direct lightning strokes are more likely to occur and where grounding conditions are not as favorable as in built-up areas. Thus, cable damage from lightning currents is more likely to result from open-wire than from drop loops. In some locations where cables extend into rural districts and are subject to the same lightning exposure as the open wire the current received by the cable from the drop loops may be of greater importance than that received from the open-wire loops connected to these cables.

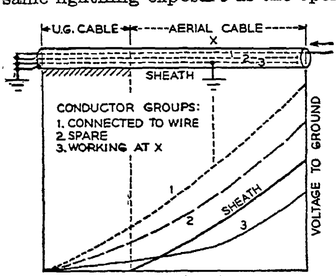


FIG. 7. Voltage Distribution along Conductors and Sheath for Current Delivered from Open Wire—No Protector Blocks on Conductors (Courtesy Bell System)

Lightning currents, reaching exchange cable from any source, produce potential differences between conductors and between the sheath and conductors which, if protection is not provided, may seriously damage the sheath and the enclosed conductors.

The potential differences within the cable, resulting from current on the conductors, depend largely on the current wave form. For steep wave-front surges, the relatively high self and mutual surge impedances are important in determining the potential differences; for slower surges or short cables (about 1 to 2 miles) these potential differences depend more upon the conductor resistances. Frequently, the slower surges produce the lower potential differences. Figure 7 illustrates the voltage distribution along conductors and sheath for current delivered from an open-wire loop with no protector blocks at the junction of the open wire and cable. The aerial cable is not grounded, except through the underground section and the office ground, and contains a group of conductors (1) connected to open-wire loops, (2) spare, and (3) working at an intermediate point, x, along the cable.

Potential differences in cables, unless several times the breakdown voltage of the cable insulation, do not, in general, cause permanent damage, although the insulation may be punctured in numerous places. Remedial measures are, therefore, designed to reduce the applied voltages between conductors and between conductors and the sheath and to limit or distribute the current so that permanent failures are unlikely even though punctures do occur. Such measures may consist of (1) connecting protector blocks between conductors and sheath where potential differences are most likely to be high and occur frequently, (2) providing means of diverting from the cable a substantial part of the current which might reach the cable over open-wire or drop loops serving installations exposed to lightning strokes, such as radio and fire towers, and (3) placing conductors parallel with the sheath, thus increasing the conductivity and reducing the *IR* drop along the sheath.

Protector Blocks. One type of protector block, commonly employed between cable conductors and the enclosing sheath at the junction of open-wire or drop loops and cables, consists of a porcelain block with a small carbon block insert held in place by a quick-melting glass cement. The porcelain block is held against a solid carbon block (which rests against a grounded metal plate) by a metal spring bearing against the small carbon insert. The insert is positioned to provide an air gap between it and the grounded carbon block of about 0.006 in. One set of these blocks is provided for each of the two conductors of the cable pair. When the current in the air gap is sufficient to melt the cement, the metal spring, which is connected to a cable conductor, forces the carbon block insert against the ground block, thus grounding the loop wire and cable conductor to which the loop wire is connected.

These protector blocks break down when steep wave front voltages of 1000 to 1500 volts are impressed, this voltage range being generally somewhat below the steep wave front voltage necessary to puncture exchange cable insulation. Protection is thus provided for the cable conductors, having protector blocks, against dielectric failure near to the point where these blocks are connected. However, damaging potential differences may occur between the protected conductors and other conductors at the point of protection and between conductors and sheath at other points. Lower-breakdown protector blocks for cable protection do not appear to offer an important improvement in protection and generally react unfavorably from the standpoint of maintenance and service.

Protector blocks applied to working conductors at the junction of open-wire and aerial cables (not grounded at intermediate points) reduce potential differences between such conductors and the sheath and other conductors, as shown by a comparison of Figs. 7 and 8. The curves in these figures are not drawn to scale, and actually the reductions obtained are much larger than can be indicated conveniently. With working conductors thus protected (Fig. 8), the maximum potential difference within the cable usually occurs at the open-wire junction between the sheath and spare pairs, and the next highest potential difference appears at the junction of the aerial and underground cable between the sheath and protected conductors.

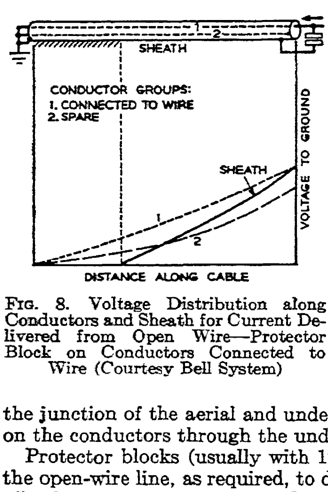


FIG. 8. Voltage Distribution along Conductors and Sheath for Current Delivered from Open Wire—Protector Block on Conductors Connected to Wire (Courtesy Bell System)

When protector blocks are applied to all conductors at the junction of open-wire and aerial cables, the potential differences between conductors are eliminated throughout the cable (assuming an ideal condition of no grounds on the conductors through station blocks at other points). The only potential difference existing in the cable under this condition is between the sheath and all conductors, being a maximum at the junction of the aerial and underground cable sections and about equal to the *IR* drop on the conductors through the underground section.

Protector blocks (usually with 10-mil air gaps) for all wires may also be placed along the open-wire line, as required, to divert current from the cable. This results in increased effectiveness for a given ground impedance, due to increasing the impedance between the ground point and the underground cable, over what it would be for a single sheath ground at the open-wire junction, and it avoids the adverse effects of a single ground on the sheath. Experience indicates that improvements usually result when the open-wire protection is placed within about 700 to 3000 ft of the cable terminal and the ground resistance at the protection does not exceed about 30 ohms.

Grounds may consist of several hundred feet of buried wire in high-resistivity areas, or rods may be used effectively in low-resistivity areas. Grounds of limited extent have impedances about equal to their d-c resistances, but for lengths of wire or pipe greater than about 200 ft the d-c resistance decreases with length faster than the impedance. Figure 9 shows, for homogeneous earth, the variation of resistance with length of buried wire, the earth resistivity being 100 meter ohms. Large variations may be expected in these values where the soil has a highly variable resistivity. For grounds of equal d-c resistances, the ground of least extent is considered best for grounding purposes, although, practically, the neutral of a well-grounded common neutral power system or water piping system, if available, will probably provide a better ground than a made ground.

Shielding conductors placed on the pole line along with the cable or buried below it will reduce conductor-to-conductor and sheath-to-conductor potentials by about 50 per cent or more. The effectiveness of shielding depends upon diverting part of the current from the sheath. This method of current diversion has been limited because more practicable methods are available.

Increasing sheath conductivity by placing bare conductors along the sheath and in contact with it will reduce the IR drop along the sheath and potential differences within the cable.

Other remedial measures include:

(*a*) Increasing core to sheath dielectric strength. This method is of major importance in toll cables, where the voltages between core and sheath appear most frequently, but for exchange cable, where the voltages between conductors are of more concern, this method is of lesser importance.

(*b*) Employing pole protection wires on exchange cables extending into rural areas where cable terminals are infrequent.

(*c*) Bonding to power system common neutrals at frequent intervals in built-up areas where many working drops are connected through protector blocks to the cable sheath. In such cases, the cable sheath becomes closely tied in with the neutral through the drops, station protector grounds, and power secondary services. By proper bonding (at about 1/4-mile intervals or less) between sheath and neutral (assuming the neutral to be continuous and well grounded), the neutral forms a parallel path with the cable for lightning currents and provides effective shielding. Under such conditions, cable damage from power contacts or lightning tends to be restricted to a section between the two bonds immediately adjacent to the power source. Bonding to the neutral also lowers the sheath impedance and thus assists in prompt de-energization of the power circuit in the event of power-circuit contact with the sheath. Frequent bonding to the neutral is useful for cable noise mitigation.

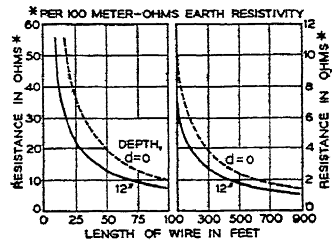


FIG. 9. Approximate Resistance to Ground of a Buried Wire (Courtesy Bell System)

Where bonding between the aerial cable and power neutral is objectionable, with respect to corrosion on the associated underground cable plant (as where drainage is used), the bonded aerial sheath may be isolated from the underground sheath by installing an insulating joint at their junction, reliance being placed in the bonding for the aerial sheath protective grounding.

Experience indicates that it is advisable, in lightning affected areas, to protect all working conductors at any terminal, serving any open wire or drop loop over 1/2 mile in length, this length being based principally on judgment.

Some companies (not Bell System) employ, in some cases, fuses (usually 5- or 7-amp rating) in conjunction with protector blocks to form a unit-type protector which is usually assembled in multiples and mounted in a terminal housing for installation at the junction of open-wire or drop loops and the cable. The protector blocks ground the open-wire or drop loop when lightning or excessive power circuit potentials are applied, and the fuses open the circuit when the current coming into the cable over the loop exceeds the fuse rating.

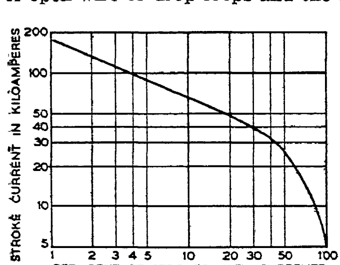


FIG. 10. Distribution of Lightning Stroke Crest Currents (Courtesy Bell System)

TOLL CABLE PROTECTION (against lightning) requirements are similar to those of the exchange plant except that the distances are greater, particularly between grounds. Experience indicates that the rate of lightning troubles on *aerial toll cables* apparently does not differ greatly from such troubles in *buried cables*, but the ease of locating and clearing these troubles in aerial cables as compared to underground cables indicates less need for comparable remedial measures on the aerial cables.

Figure 10 shows the distribution of lightning stroke crest currents, based on a large number of measurements in the ground structures of power-transmission lines, the crest value varying over a wide range. Measurements indicate that the crest value is reached in 5 to 10 microseconds and that it decays to half its maximum value in 25 to 100 microseconds. The crest value and decay time to half value of the current, rather than wave front steepness, are of primary importance with respect to voltages between the sheath and conductors of aerial or buried toll cable.

At a point remote from ground connections, the surge impedance to ground of an aerial cable sheath (any size) is about 200 ohms (400 ohms in each direction from the lightning stroke). For a crest current of 20,000 amp the sheath-to-ground voltage would be 4 million volts. The conductors within the sheath attain about the same potential to ground as the sheath, owing to capacitive and inductive coupling between the sheath and core conductors. However, some voltage difference does exist between the sheath and core, due to the IR drop along the sheath, resulting from the flow of lightning current on the sheath. This voltage difference is greatest at the stroke point, decreasing as the distance from the stroke point increases. The higher the sheath resistance, the greater this voltage difference between sheath and core will be for a given lightning current.

The dielectric strength between core and sheath, for normal-dielectric-strength cables, is for surges about 2000 volts. For high-dielectric-strength cables (having an extra core wrap), a 4000-volt value for surges may be assumed.

After an initial puncture in the insulation near the stroke point, other punctures will thus usually occur some distance away in either or both directions. Such puncturing may or may not cause permanent failures, depending on the current through the fault.

A sheath-to-core voltage of breakdown magnitude may develop, owing to: (1) a large sheath current for a relatively short distance, (2) a smaller sheath current over a longer distance, or (3) a combination of (1) and (2), depending on the ground paths from the sheath and their locations with respect to the lightning stroke. Permanent damage is likely to result from the large current transfer through the punctures in condition (1) but is not so likely in (2) because of the smaller transfer of current from sheath to conductors under this condition.

The voltages in buried cables are usually due to large sheath currents over short distances (less than 1/2 mile), so that permanent damage is likely at puncture points.

The protection of aerial toll cables may consist of (1) pole protection wires (wires bonded to the sheath and extending down the pole to the pole butt or to a point near the ground line), (2) aerial shield wires, (3) buried shield wires, (4) high-dielectric-strength cable (with double core wrap), (5) shields within the cable, or (6) protector blocks at open-wire junctions or out from the junction about a mile on the open wire.

Pole protection wires may serve not only to protect poles against splintering but also to provide protection to aerial cables in lightning exposures. Data show that for such wires, properly spaced along an aerial cable line, the voltage between sheath and core, due to a stroke, decreases as the number of such wires increases up to about 10 for 100 meter ohm resistivity and up to about 20 for 1000 ohm resistivity. However, in uniform exposures, the effectiveness of these wires decreases for aerial cables having both toll and exchange conductors as the interval (of the order of 1 mile) between subscriber drop locations decreases, owing to the low-breakdown path from sheath to ground over the drop loops which substantially short-circuits the pole protection wires.

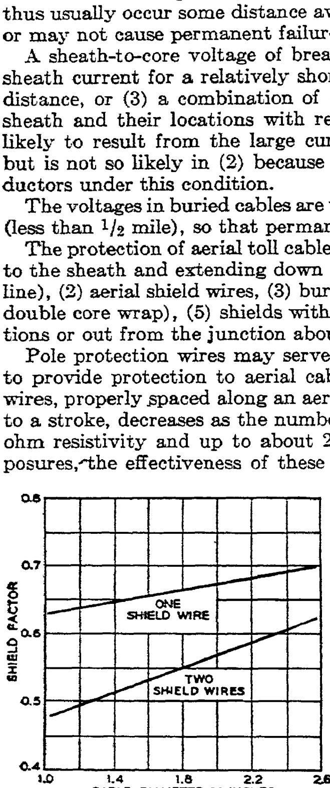


Fig. 11. Shield Factors for Aerial Shield Wires (Courtesy Bell System)

The effectiveness of these wires may be lowered where a large number of them are employed, as the result of current distribution along the sheath to such grounds from the stroke point increasing the net sheath-to-core voltage. By using proper length gaps in the wires, the number of conducting grounds may be limited to an optimum value.

Where pole protection wires are not adequate to give required lightning protection to aerial cables, aerial or buried shield wires may be of advantage. The reduction in core-to-sheath voltage by aerial shield wires which are bonded to the sheath at each pole is normally substantial, as shown by the shield factors, Fig. 11. When an aerial cable is provided with buried shield wires, the voltage from a direct stroke is about the same as that for a buried cable of the same size with similar shield wires.

High-dielectric-strength cables should generally be employed in new installations. The choice of this type of cable may obviate the necessity of other remedial measures, but in any event its higher dielectric strength is an advantage since, if the strength is doubled, the stroke current must be increased 3 times to puncture the insulation.

Communication plant may be damaged by contact with high-voltage wires or power distribution circuits. The most important consideration in protecting cables from power-wire contact is to so locate, construct, and maintain the communication and power circuit plant that contacts will not occur.

Aerial cables are grounded at offices, through their connection to the underground cable plant, underground dips, and private cable entrances, as well as other frequent grounding, such as, in some cases, to multigrounded power circuit neutrals. This practice contributes to provide a low-impedance ground to de-energize the power circuit promptly in case of contact with the cable plant.

27. CABLE SHEATH CORROSION

ELECTROLYSIS. Electric currents, flowing in the earth, may result from (1) stray currents from d-c street-car (trolley) or electrified railway systems, (2) stray currents from commercial d-c power distribution systems, (3) differences in the chemical composition of films on the cable sheath at different locations, (4) differences between the composition of films on sheaths of different cables at the same location, (5) differences in the electrolyte at different places, or (6) a number of other conditions. These currents passing through damp ground cause chemical changes to take place at electrode surfaces, such as cable sheaths, which may or may not affect the electrodes.

Cable sheath corrosion may, in general, be considered as occurring in:

1. Stray-current areas.
2. Non-stray-current areas.

The principal causes of cable sheath corrosion are:

1. Stray current-anodic action.
2. Stray current-cathodic action.
3. Localized action.
4. Galvanic action.
5. Chemical attack.

Stray-current areas are commonly designated as those in which the cable-to-earth potentials and sheath currents are established principally by currents straying from the rails of d-c transportation systems and utilizing other paralleling paths of relatively low resistance in the earth, such as cable sheath or public water piping.

Where the stray current enters the sheath the cable is negative (cathodic) to the earth, and where it leaves the sheath the cable is positive (anodic) to the earth.

Non-stray-current areas are those in which corrosion may occur from other than stray currents. Currents in these areas usually result from potential gradients due to such causes as differential aeration, differential electrolyte concentrations, non-uniformities in the sheath metal, or potential differences between different metals.

Current from stray-current areas may flow in non-stray areas, such as along an underground toll cable sheath extending between cities where street cars operate. Current, usually considered as "non-stray" current, such as that resulting from galvanic potentials, may also be present in stray-current areas, although its effect is largely overshadowed by the stray currents in these areas.

Anodic corrosion, due to stray currents, results because:

1. The metal becomes positively ionized and goes into solution.

$$\mathrm{Pb} - 2e \rightarrow \mathrm{Pb}^{++} \quad (e = \text{electron})$$

2. The anions (ions bearing negative charges and migrating to an anode) in the electrolyte are attracted to and contact the sheath, where they lose their charge, become chemically active, and attack the lead sheath.

If the ions are chlorine

$$\mathrm{Cl}^{-} - e \rightarrow \mathrm{Cl}$$

and

$$\mathrm{Pb} + 2\mathrm{Cl} \rightarrow \mathrm{PbCl_2} \quad \text{(lead chloride)}$$

Note: A positive ion, in the case of Pb^{++}, is an atom from which two electrons have been removed; a negative ion, in the case of Cl^{-}, is an atom to which one extra electron has been added.

In anodic action, a corrosion product does not always adhere to the sheath, and a clean corroded area on the sheath is regarded as due to such action. The corrosion product, if present on the sheath, is usually lead chloride or lead sulfate or both. The sheath potential, being positive, attracts chloride and sulfate ions in the ground water to the sheath, and these ions react with the lead. Under severe anodic conditions lead peroxide, PbO_2, may be formed. This product has a chocolate color; the products of chloride, sulfate, and carbonate of lead are white.

Cathodic corrosion, due to stray currents, though not as yet reproduced in the laboratory, might possibly occur when the cables are negative to earth and the surrounding earth contains alkali or salts of sodium, potassium, calcium, or magnesium. Hydrogen ions in the electrolyte, under such conditions, are attracted to the sheath, lose their charge, and are liberated, resulting in a decrease in hydrogen-ion concentration and making the electrolyte alkaline. The alkali forms most rapidly and with greatest concentration at points of maximum cathodic current density. The attack on the sheath is, therefore, not uniform and usually affects only a relatively small part of the sheath.

In the case of sodium salts, the reaction with the lead is:

$$H^+ + e \rightarrow H$$

$$Na^+ + OH^- \rightarrow NaOH \quad \text{(sodium hydroxide)}$$

The sodium hydroxide forming at the sheath dissolves the lead, and the final reaction usually results in the formation principally of *lead monoxide*, PbO, lead carbonate, and sodium carbonate. The sheath pitting may be similar to that caused by anodic action, but the cathodic corrosion is characterized by the usually bright orange-red color of the lead monoxide.

Galvanic action usually results when two dissimilar metals are in electric contact with an electrolyte and also are metallically connected. Where cables and bare copper or galvanized cable rack supports are present in a flooded manhole, galvanic action may occur. Lead is anodic to copper and cathodic to the zinc galvanizing, the anode corroding in each case. Corrosion may appear, in the presence of moisture, on wiped joints and soldered seams on cable sleeves, where the solder may be anodic to the lead.

When two dissimilar metals in an electrolyte are connected metallically, the metal with the higher solution pressure becomes the anode and corrodes. Current passes from the anode through the electrolyte to the cathode and completes the circuit to the anode through the metallic connection.

In alkaline solutions, lead tends to become electronegative and may be anodic with respect to another metal such as iron. Thus, lead corrosion may result where cable is installed in iron conduit and such solutions are present. Under these conditions current may pass from the lead through the solution to the iron and return to the lead at points of metallic contact between the iron conduit and the sheath.

Local action between adjacent areas of the sheath may result from variations in sheath material, such as foreign particles in the lead, and from differences in surface conditions due to abrasion. Corrosion from the action, under anodic conditions, may occur in spots rather than uniformly over the sheath surface, and if the action continues for a period of years sharply defined pits may appear in the sheath.

Concentration cells may be formed by a change in the concentration of the salts in an electrolyte, causing corresponding changes in the potential of a given electrode in contact with the electrolyte; or by two electrolytes, with equal concentration of different salts, producing different electrode potentials on the same metal. Such a cell might be established by waste products from a factory or sewer entering a sloping conduit (with cable) at a conduit joint and concentrating along a section of sheath. The cell might then result between the section of sheath having the concentration and an adjacent section not affected by the waste.

Differential aeration, which may be considered a special form of concentration cell, results in sheath corrosion, owing to a variation in the concentration of dissolved oxygen in the electrolyte in contact with the sheath. Where a sheath is continually wet by water dripping or condensing on it, there will be more oxygen in the moisture exposed to the air than where the moisture is shielded by an absorbent material such as a layer of silt on the sheath. The cable sheath in contact with the electrolyte with a deficiency of oxygen will be anodic and subject to corrosion.

Battery action in the soil might result in sheath corrosion where the sheath was not one of the cell electrodes but acted as a conductor of the local current. As an example, if an iron pipe passed through a bed of cinders which acted as a carbon electrode, and the adjacent section of pipe was in contact with ordinary soil, a cell would be established with the carbon electrode negative to the adjacent section of pipe. Part of the resulting current from this section might enter a nearby paralleling cable sheath and leave the sheath near the cinders, at which point sheath corrosion might result.

Protective films formed over the sheath by natural processes are usually helpful in preventing sheath corrosion. Films formed from silicates are generally continuous, adherent, insoluble, and helpful, whereas films resulting from nitrates in the soil tend to prevent film formations and thus aid corrosion.

REMEDIAL MEASURES for controlling corrosion are varied and often complex, because the problems of corrosion may differ over a wide range of causes and conditions and extend over relatively large areas.

Generally, the first considerations in the control of cable sheath corrosion are electrical, involving a thorough study of electrical conditions affecting the cable plant. Chemical control can sometimes be used successfully where electrical methods are not practicable. The general attack on sheath corrosion problems by electrical methods consists of:

1. Limiting current entering the sheath to the extent practicable.
2. Providing metallic paths by which current entering the sheath may leave it without damage to the sheath or other metallic structures.

Protective arrangements of any nature should be such as to minimize the probability of impressing current on other underground plants, either privately or publicly owned.

The limitation of current pick-up by a given underground cable sheath requires that the sheath be kept free of all connections to other grounded metallic structures except those connections specified as part of the general remedial plan, and that the sheath should not be made more negative to earth in any area than is necessary under practical design.

The cable sheath should be maintained at slight negative potential to earth to (1) limit current pick-up which increases with increase in negative potential, (2) lessen the possibility of large positive pipe-to-cable voltages, and (3) reduce the chances for cathodic corrosion.

The increasing use of bonding between aerial cables and multigrounded power neutrals for noise induction and protection reasons increases the tendency to discharge current to the underground cable plant. Insulating joints may be used at underground-aerial cable junctions, where required, and electrolytic capacitors may be employed to bridge the insulating joint if capacitors are necessary for noise induction or protection purposes.

Cathodic corrosion of sheaths may occur where the cable-to-earth potential exceeds a few tenths of a volt negative. In the presence of salts this potential may not be over 0.2 volt; under other conditions cathodic corrosion will not result at any potentials encountered, where an effective drainage system exists.

For this type of corrosion, it is essential to maintain low negative cable-to-earth potentials by using (1) an adequate drainage system, (2) insulating joints shunted by resistors if necessary, (3) corrosion-protected cables for replacement of cables which fail, (4) periodic flushing of ducts which accumulate alkali, or (5) reverse drainage employing controlled currents *to* the cables by connecting to positive points on rail systems or to pipes (experience limited).

Other underground structures, such as public water or gas piping, are also drained in many areas. A metallic piping system in direct contact with the earth usually has a very low leakage resistance, and any reduction in its potential by drainage tends to lower the earth potential in its immediate vicinity. Where the earth potential is thus lowered, the current discharge from nearby telephone cables tends to increase. This condition may complicate the telephone cable drainage scheme, but generally, owing to the low leakage resistance of the piping system, the adverse effects are not extensive. A coordinated drainage plan for all underground systems is usually necessary where two or more systems serve the same general area affected by stray currents.

Anodic corrosion, which is the only type of corrosion that has been found so far in non-stray areas, results from current leaving the cable sheath for any reason. This type of corrosion may be due to: (1) currents flowing from the cable into the electrolyte in the duct and back to the cable without leaving the duct, as a result of differential aeration, differential electrolyte concentration, or non-uniformity of the sheath composition (though this type of corrosion is local, many such *local cells* may exist along a cable and cause corrosion over a long section of the cable); and (2) currents flowing from the cable to the electrolyte in the duct, thence to the earth outside of the duct, and finally returning to the cable at some relatively remote point. In (2) the driving electromotive forces may result from the same causes as given in (1) above, but the cells are materially lengthened and the currents are known as *long cell currents*. Such electromotive forces may result from potential gradients in the earth due to currents associated with corrosion cells on long paralleling piping or to other natural causes, including magnetic storms; or they may result from a potential established as a result of sheath contact with other metal, such as copper or iron piping; or they may result from remote railway currents flowing into the non-stray area.

TESTING METHODS AND MITIGATIVE MEASURES, as employed in non-stray-current areas, are, in general, somewhat similar to those employed in stray-current areas, but various testing refinements are usually required in the non-stray areas. For example, in determining the *IR* drop or direction of current flow in the earth as it may affect underground cable sheath, it is necessary to employ electrodes, for contacting the earth, which are identical in potential or differ by a known and constant potential. Such an electrode, designated a *half-cell*, consists of a non-conducting container enclosing a metallic electrode suspended in an electrolyte. The electrolyte fills a porous cap, which forms the bottom of the container and through which the cell makes contact with the earth. This cell makes use of the fact that the potential difference between electrolytes varies over a very small range as compared to the potential difference between metals when used as electrodes.

Experience has shown that corrosion is not usually a problem with respect to buried jute-protected, tape-armored, and thermoplastic-covered cables in non-stray areas. Electrical tests on such cables in non-stray areas are not, in general, considered necessary.

Forced drainage has been successfully employed in anodic areas in the protection of underground cable plant in non-stray areas. Random contacts between telephone cables and other metallic structures, such as water piping and steel buildings, usually interfere with effective drainage of the cables and should be eliminated, although other considerations, particularly noise induction and protection, may impose numerous difficulties and problems in such eliminations.

Forced drainage requires a separate d-c source of potential, such as a rectifier (commonly used) or a battery, with the negative terminal connected to the cable sheath and the positive terminal connected to a negative bus or special ground. Current is forced from the sheath to the bus or ground.

Rectifiers available for drainage purposes have suitable d-c voltage and current outputs and usually operate from either 115- or 230-volt a-c commercial supply. They are made in several types, including dry-disk and tube types, by a number of electrical equipment manufacturers.

Galvanic anodes requiring no external power supply may provide the required amount of forced drainage under favorable conditions when buried in the ground and connected to the sheath by copper wire. The anodes may consist of a metal negative to lead, such as zinc, aluminum, or magnesium. Magnesium appears favorable, because of its relatively high negative potential (about 1 volt) to lead. The anode, being buried in the soil and discharging current, will gradually be consumed.

To be effective, the anode must have a low resistance to ground, and this depends on its shape and the surrounding earth resistivity. For a cylinder 4 in. in diameter by 20 in. long, the resistance to ground (without special environment) is about equal to the earth resistivity in meter ohms.

Chemical attack usually requires an analysis of the corrosion products and a determination of the source of the chemical attacking the cable sheath in order to apply suitable remedial measures. Chemists may be of assistance, in difficult cases, in determining the nature of the attacking chemicals. Lead monoxide corrosion is indicative of cathodic action or alkali attack. Very few cases show a definite single cause of corrosion. The previous history of corrosion of the affected cable may be of value in arriving at the causes of the corrosion.

Alternating current, principally because of its rapid and equal reversals of potential between positive and negative values, is not considered an important cause of cable sheath corrosion.

Corrosion-protected cable for installation in underground conduit can be made available with the same core make-up as the plain lead-covered cable with which it may be associated. This protected cable may be useful where the lead sheath would be subject to corrosive action without the protection and where it is more attractive than other remedial measures. This type of cable may be employed in situations such as (1) near chemical plants or other locations where chemical attack has been experienced or may occur, (2) where alkaline attack (cathodic action) might develop, or (3) where corrosion has occurred in subsurface dips.

One type of protection consists of two reversed layers of Sisalkraft paper and an outer layer of rubber-filled tape, the sheath and each layer being flooded with an asphalt compound. A non-adhesive coating is applied on the outside covering to prevent sticking in handling. The protection increases the cable diameter about 0.2 in.

Suitable alarms and pilot wires to a centralized maintenance center may be employed, when facilities are available and as required, to indicate critical changes in potentials and current flow affecting the cable plant.

COORDINATION OF COMMUNICATION AND POWER SYSTEMS

By John D. Taylor and Howard L. Davis, Jr.

28. FOREIGN WIRE RELATIONS

In order to insure safety to persons and property, economy of operation, and good service, in areas served by both overhead communication and power systems, it became evident, as the systems began to expand, that the companies involved should establish and follow a plan of cooperation in the construction and operation of their respective plants. For a number of years individual cooperative efforts were carried on, but the specific solutions of the problems that developed were not applicable in a general way.

Early in 1921 steps were taken by both interests on a nationwide scale to formulate a basis of common understanding and to establish permanent joint committees and subcommittees for study and recommendations relating to mutual problems. As a result of continuous study and research by these subcommittees and sponsors, the Joint General Committee of the Edison Electric Institute and Bell Telephone System prepared, and the representative interests approved, several general reports, of which the following are the principal ones in effect today:

1. *Principles and Practices for the Inductive Coordination of Supply and Communication Systems*, Dec. 9, 1922.*
2. *Principles and Practices for the Joint Use of Wood Poles by Supply and Communication Companies*, Feb. 15, 1926.*
3. *Inductive Coordination—Allocation of Costs between Supply and Communication Companies*, Oct. 15, 1926.*

In general, the *Principles and Practices* provide, in addition to other important items, that

(*a*) All supply and communication circuits with their associated apparatus should be located, constructed, operated, and maintained in conformity with general coordinated methods based on the concept of rendering either service without interference.

(*b*) Where general coordinated methods will be insufficient, suitable specific coordinated methods should be applied, most conveniently and economically, to prevent interference with either service, present and known future factors being taken into account.

(*c*) The companies serving any given area should fully cooperate with each other in carrying out the accepted principles, based on arriving at the best engineering solution of each situation, as it arises, for all the companies involved.

(*d*) Where conditions and the nature of the supply and communication circuits permit, joint use of poles (particularly in urban areas) is generally preferable to separate lines, when justified by considerations of safety, economy, and convenience, and assuming that a satisfactory agreement is reached between the parties concerned.

(*e*) When supply and communication facilities occupy the same section of highway and joint use is not desirable, each type of facility should be confined to one side of the highway, as far as practicable, thus avoiding unnecessary crossings and expensive guying.

(*f*) In the design, construction, and operation of supply and communication circuits and equipment, all factors contributing to inductive influence, inductive couplings, or inductive susceptiveness under normal or abnormal operating conditions should be limited to the extent necessary and practicable.

(*g*) Each utility shall be the judge of the quality and requirements of its own service and the type and design of its own facilities.

(*h*) Coordination costs in any given situation of proximity, assuming that satisfactory results have been attained under the best engineering solution, will generally be allocated, so that each company involved bears its equitable portion, including its own betterments.

The basis for cooperation, as set forth in detail in the *Principles and Practices*, has contributed immeasurably to the excellent foreign wire relations existing among the many wire-using companies who supply electric and communication services throughout the country.

The cooperative plan was later extended to provide for a Joint General Committee of the American Railway Association (now Association of American Railroads) and Bell Telephone System in 1929, a Joint General Committee of the Edison Electric Institute

---

* These reports are now combined as *Reports of Joint General Committee of Edison Electric Institute and Bell Telephone System on Physical Relations between Electrical Supply and Communication Systems*, reissued July 1945.

and Western Union Telegraph Company in 1935, and a Joint General Committee of the Association of American Railroads and Edison Electric Institute.

When power and communication facilities serve the same areas and are supported on overhead structures and where the two types of facilities are in close physical relation or inductively coupled or both, situations of proximity generally are unavoidable. Communication facilities, operating at relatively low voltages and currents, are not designed to withstand the normal or abnormal power circuit voltages and currents which may be impressed upon them by direct contact or, in severe influences or couplings, by induction. In some cases of contact or induction, the service may be only interrupted or degraded, but in more severe situations the service may be interrupted and the facilities damaged with or without personnel hazard.

It is thus imperative that the *Principles and Practices* be closely adhered to, in order that both services may be furnished the public in a safe, economical, and satisfactory manner.

The design and application of coordinative measures by the wire-using companies involve two broad fields of coordination, structural and inductive. Both fields have been for a number of years and will continue to be under intensive study, directed toward improving methods and securing further economies without sacrificing but, when practicable, bettering present safety and service.

29. STRUCTURAL COORDINATION

Structural coordination, as the name implies, consists of planning, designing, constructing, and maintaining the physical overhead plant of each company, with due consideration for the plans and plant of all other companies involved, so that safety and overall economy will be attained.

SITUATIONS OF PROXIMITY are created when power and communication lines are so located with respect to each other as to parallel, cross-over, occupy the same poles (joint use), or require consideration of line wire, guy, or pole-mounted equipment clearances. The overbuilding of one type of service by the other is considered objectionable, joint use usually being the better solution, with the power wires in the upper position in all cases of proximity.

The National Electrical Safety Code, Fifth Edition, Part 2 (*National Bureau of Standards Handbook* H32), issued Sept. 23, 1941 (hereinafter referred to as the Code), which has been approved by the American Standards Association, presents Safety Rules for the Installation and Maintenance of Electric Supply and Communication Lines. These rules embody specific minimum requirements, and, though not complete, they are intended to cover those points which are most important for the safety of employees and the public. This code is acceptable to the various wire-using organizations throughout the country and, except as modified by more exacting state or local regulations, is applied, together with the *Principles and Practices* discussed previously, generally throughout the country. Where a certain coordinative problem arises, not specifically covered in the Code, the companies involved agree on the best engineering solution for the problem. The Code also provides for modifying or waiving its requirements in any given case where such requirements are inapplicable, not justified, or impracticable, or where equivalent or safer construction can be more readily provided by other means. It is, therefore, a practicable and flexible guide.

Structural requirements and clearances between electric, railway, and communication structures, wires, and equipment must be adequately provided for, as set forth in the Code. Owing to the large amount of detailed information necessary in specifying these requirements under the numerous conditions encountered in practice, they will not be given here, but they may readily be obtained from the Code.

Certain fundamental concepts (other than those previously enumerated under *Principles and Practices*) in this work are generally accepted as good engineering practice and may be stated in general terms:

(*a*) The mechanical design and construction of electric (supply) and communication systems should conform to good modern practice.

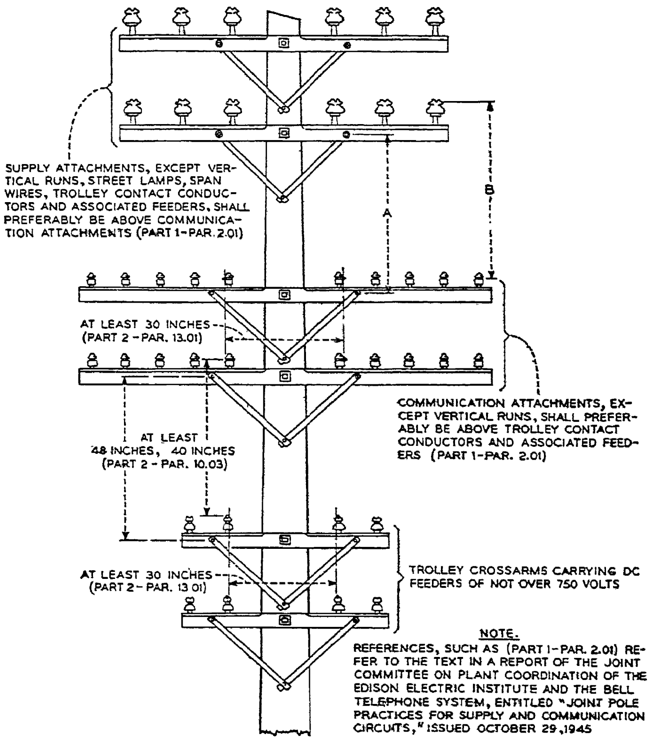
(*b*) When changes are made in systems or methods of operation, consideration should be given to decreasing inductive influences and susceptiveness, when practicable.

(*c*) Coordinated systems should be maintained, so that abnormal conditions affecting either service will be minimized and prompt action will be taken to eliminate such conditions when they do occur.

(*d*) Supply, communication, and trolley circuits should occupy levels in the order named, with supply circuits at the top level. Also, where supply lines carry different voltages, the higher-voltage lines are usually placed above those of lower voltage.

(*e*) Joint use should be considered when it can be employed with reasonable safety and convenience, economically, and without appreciable service detriment.

JOINT USE of poles is a very desirable means of coordinating supply and communication facilities where this type of construction is feasible. Various types of construction are employed, depending upon the types of facilities involved, but in any case the construction is designed to conform to the latest practices and safeguards.



VOLTAGE OF SUPPLY CIRCUIT CONCERNED	DIMENSION	MINIMUM CROSSARM SPACING IN INCHES	DIMENSION	MINIMUM SEPARATION BETWEEN CONDUCTORS IN INCHES
0-8,700	A	48	B	40
OVER 8,700		72		60

(PART 2- PAR. 10.03)

FIG. 1. Relative Position of Attachments, Showing Vertical Clearances and Climbing Space (Joint Pole Practices for Supply and Communication Circuits, Oct. 29, 1945)

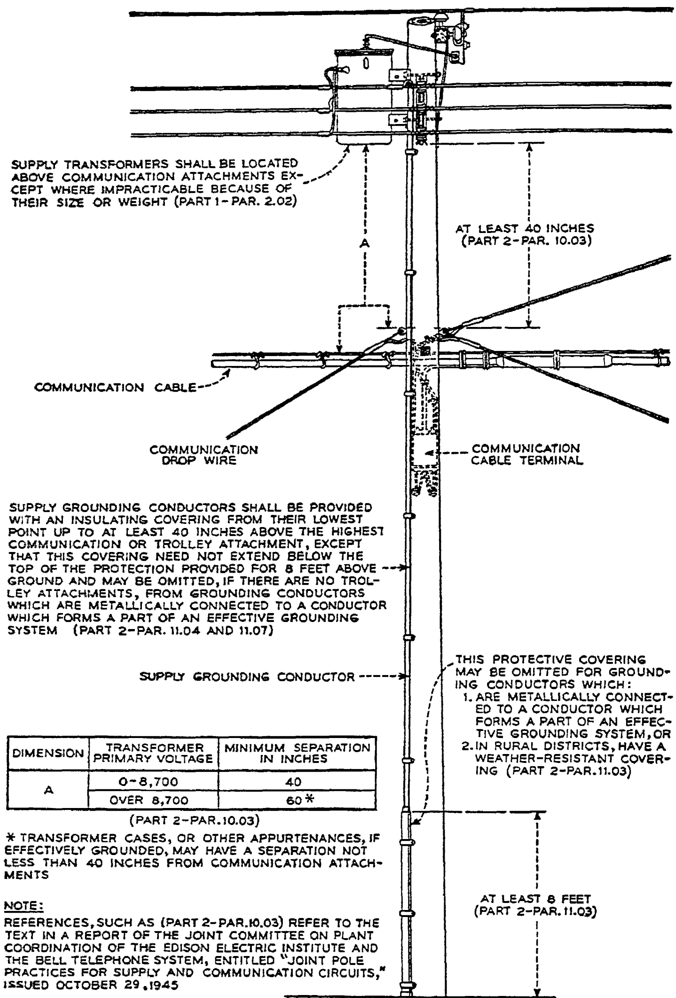
For joint-use construction, commonly employed in urban areas for local distribution, Figs. 1, 2, 3, and 4 show the usual typical construction features and clearances considered good modern practice and meeting Code requirements. Adequate clearances are essential for the protection of personnel and property.

Normal joint-use construction is applicable to construction involving communication cables or conductors and supply cables or conductors of the following types:

(*a*) Constant-potential a-c supply circuits normally operating at voltages between 750 and 5000 volts between conductors and not over 2900 volts to neutral or ground.

(*b*) Constant-current supply circuits of not more than 7.5 amp regardless of the voltage, and of more than 7.5 amp where the open-circuit voltage of the supply transformer is not more than 2900 volts.

(*c*) Constant-potential a-c supply circuits normally operating at more than 5000 volts between conductors or more than 2900 volts to neutral or ground, and constant-current



DIMENSION	TRANSFORMER PRIMARY VOLTAGE	MINIMUM SEPARATION IN INCHES
A	0-8,700	40
	OVER 8,700	60 *

(PART 2-PAR. 10.03)

* TRANSFORMER CASES, OR OTHER APPURTENANCES, IF EFFECTIVELY GROUNDED, MAY HAVE A SEPARATION NOT LESS THAN 40 INCHES FROM COMMUNICATION ATTACHMENTS

NOTE:
REFERENCES, SUCH AS (PART 2-PAR. 10.03) REFER TO THE TEXT IN A REPORT OF THE JOINT COMMITTEE ON PLANT COORDINATION OF THE EDISON ELECTRIC INSTITUTE AND THE BELL TELEPHONE SYSTEM, ENTITLED "JOINT POLE PRACTICES FOR SUPPLY AND COMMUNICATION CIRCUITS," ISSUED OCTOBER 29, 1945

FIG. 2. Supply Transformer Installation, Showing the Separation from Communication Cables and Conductors (Joint Pole Practices for Supply and Communication Circuits, Oct. 29, 1945)

supply circuits of more than 7.5 amp where the open-circuit voltage of the supply transformer is more than 2900 volts, provided that:

1. The supply and communication circuits are so constructed, operated, and maintained that the supply circuits will be promptly de-energized, both initially and following subsequent breaker operations, in the event of contact with the communication plant.

2. The voltage and current impressed on the communication plant, in the event of a contact with the supply conductors, are not in excess of the safe operating limit of the communication protective devices.

Note: Where conditions 1 and 2 above are not met, special joint-use construction is required, unless the parties concerned agree that the additional protection, required under (*c*) above, is unnecessary or except in the case of certain drop wire attachments.

(*d*) Any effectively grounded supply cables, located above communication cables or conductors or carried on effectively grounded suspension strand, where the supply voltage between conductors is more than 750 volts.

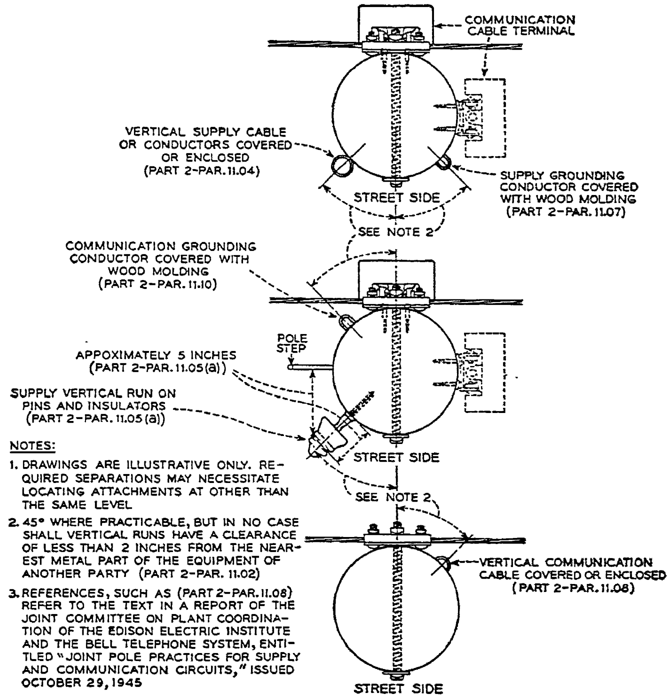


FIG. 3. Location of Vertical Runs (Joint Pole Practices for Supply and Communication Circuits, Oct. 29, 1945)

The requirements for normal and special joint use conform to Code Grade C and B construction, respectively, applying mainly to supporting structures, supply conductors, and clearances.

JOINT-USE RURAL LINE CONSTRUCTION at the higher voltages (above 5000 volts) for power and telephone services, though recognized for some time as the best engineering solution for many situations of power and telephone lines along the same route and though employed in various specific cases, has presented a number of problems in coordination. In particular, the longer spans and higher voltages for the power circuits and the increased noise induction from the longer exposures than are normally encountered in joint-use urban construction are main factors. The rapid growth, within about the last decade, of rural electrification has greatly emphasized the necessity for a study and solution of these problems, to permit a more general application of rural joint-use construction.

For the purposes of study of safe and economical rural joint-use construction at the higher voltages, several projects of this type were completed and placed in service before 1947 in Alabama, in the light-loading district, and in Minnesota and South Dakota, both in the heavy-loading district.

The NES Code (Fifth Edition of Part 2) provides that, if the supply circuit will be promptly de-energized in the event of accidental contact with the telephone plant, and the resulting voltages on the telephone plant from such a contact will be within the operating capabilities of the telephone protective equipment, Grade C construction may be employed in joint use. This provision of the Code has an important bearing on the problems involved in long-span rural joint use.

For long-span construction, high-strength line wires are necessary; they are generally of stranded copper or aluminum with one or more strands of steel for power circuits, and of high-strength solid steel or copper-covered steel for telephone circuits. The minimum size of power wire for Grade C construction under the Code is No. 8 AWG medium-hard drawn copper.

With high-strength line wires for both power and telephone conductors, it was assumed that their sag characteristics would be sufficiently alike to prevent contacts in the span under ice or wind loadings or temperature changes, with reasonable minimum separations of the two classes of conductors at the poles. It was also assumed that a very large percentage of rural power circuits would consist of one phase wire on a pin at the top of the pole and a multigrounded neutral below it on a pin or secondary rack, or that the primary circuit would consist of two or more primary wires on a cross-arm at the top of the pole and that the multigrounded neutral, if present, would be located on the same cross-arm or below it.

Where joint rural lines cross fields or other property, which is or is likely to be traversed by loaded vehicles or farm machinery, adequate wire clearances should be provided.

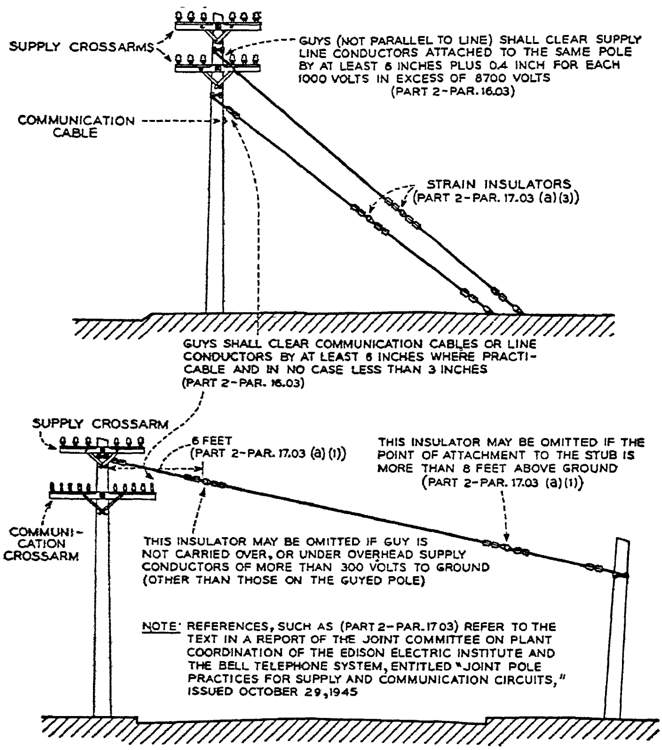


FIG. 4. Use of Strain Insulators in Ungrounded Guys (Joint Pole Practices for Supply and Communication Circuits, Oct. 29, 1945)

Span lengths in the Alabama installations of rural joint use average about 400 ft with maximum spans of 600 to 800 ft. In Minnesota and South Dakota, the loading conditions being more severe, the average and maximum span lengths are about 320 ft and 400 ft respectively.

For noise induction reasons, the separations between the power and telephone circuits should be kept as uniform as practicable without incurring undue costs or construction difficulties. (See article 31.)

Electrical protection, to meet requirements, should provide: (*a*) fuses, circuit-breakers, or other devices, which will promptly and reliably de-energize the *power circuit*, if a ground fault of relatively low impedance occurs; (*b*) protective gaps, to be connected between the *telephone circuit* and the common neutral (or other low-resistance ground) at such intervals that the impedance to ground of the telephone plant, in case of a power wire contact, will be low enough to permit proper operation of the power protective devices.

As experience indicates that the frequency of contacts is very low, the power protective arrangements are usually more than adequate to meet the above requirements. The telephone circuits in the joint projects, referred to above, employ protectors having a 3-kv (rms) breakdown to ground, placed at about $^1/_2$-mile intervals, thus providing a low-impedance path to ground. The current-carrying capacity of the protector must be sufficient to meet the discharge which may result from a contact.

As a means of limiting induced 60-cycle open-circuit voltages (under normal operation) on telephone circuits, which are occasionally disconnected from the central office equipment and entrance cables, a drainage device, consisting of a 0.25-μf condenser in series with a 10,000-ohm resistor from each wire to ground, may be used.

Although experience at present is limited, the feasibility of higher-voltage long-span joint use in rural areas has been established, and economic studies in progress indicate that worth-while economies are possible for most new extensions and sometimes even where power lines have already been built.

HIGHER-VOLTAGE CROSSINGS (involving supply lines of more than 5000 volts between wires or 2900 volts to neutral or ground) are given special consideration to minimize possible contacts between supply and communication circuits. The Code requirements for spacings, clearances, and strength of construction are considered to be minimum and represent the generally accepted practices (except as modified by more exacting state or local regulations) throughout the United States.

RURAL POWER LINE CARRIER TELEPHONE SYSTEMS, where applicable, are being placed in operation throughout the country as a means of providing telephone service to rural subscribers who are accessible from rural power-line systems but not readily so from rural telephone lines. Although these carrier telephone systems are still in their early trial periods, it appears that they will eventually be economically useful in furnishing telephone service to distant farms and ranches which would otherwise require costly telephone-line construction to reach. Joint practices and agreements covering this type of operation are in the formative stage.

A more detailed description of this carrier system and its associated installation and operating features is given in Section 17.

30. INDUCTIVE COORDINATION

Inductive coordination, as it applies to supply (electric) and communication companies, embraces the principles and practices, agreed upon by these companies, for the design, location, construction, operation, and maintenance of their respective systems, located in the same general territory, in such manner as to prevent interference with the furnishing of either the electric or communication service.

Since much of the general public throughout the country receives both electric and communication service by means of overhead construction, it is inevitable that many situations of proximity between the two services will be created under a wide variety of construction and operating conditions. Also, telephone circuits usually transmit relatively low amounts of electrical speech power (varying from less than 1 to a few milliwatts), whereas electric transmission and distribution circuits transmit power ranging from a few to hundreds and thousands of kilowatts. Owing to the much greater amounts of power carried by these circuits, not only at the fundamental but also in many cases at harmonic frequencies within the voice-frequency range, inductive effects from power circuits may, under certain conditions of proximity and operation, severely affect communication service, if proper control is not provided.

The control of inductive interference in communication circuits, to permit giving a satisfactory service, is accomplished through the application of general coordinated methods, or, when required, by applying specific coordinated methods, as set forth in the *Principles and Practices*, discussed in article 28.

There are two general classifications of inductive interference, namely: (*a*) noise frequency (induction from power harmonics within the voice-frequency range); (*b*) low-frequency (induction from the power fundamental frequency, usually 60 cycles, during abnormal power circuit conditions).

Inductive effects between power and communication circuits arise from the fact that power wires, transmitting relatively large a-c voltages and currents, establish strong electric and magnetic fields in their vicinity, which, owing to their varying character, set up in nearby communication wires alternately increasing and decreasing electric voltages and currents of the same frequencies as those in the power wires. Induction from communication to power wires would have no noticeable effect in any event on power service, because of the small amount of power carried by the communication wires and the nature of the power service.

It is often desirable to consider effects of magnetic and electric induction separately, particularly in the technical analyses of specific problems. This is not only because the physical processes and the effects of voltage and current induction are quite different but also because the power-circuit voltages and currents are often affected differently by changes in conditions. Electric induction is due to the voltages on the power line; magnetic induction is due to currents on the power line.

Theoretically, electric and magnetic induction are produced as described briefly below.

A simple method of visualizing *electric induction* is by means of the capacitances involved with a single power wire and a single telephone wire, as shown in Fig. 5. Neglecting the impedances outside the exposure (shown dotted in Fig. 5), the voltage of the power wire to ground (E_P) divides over the capacitances C_{TP} and C_{T0} in proportion to their impedances (that is, in inverse ratio to their capacitances). The induced voltage (E_T) on the telephone wire may therefore be expressed mathematically as:

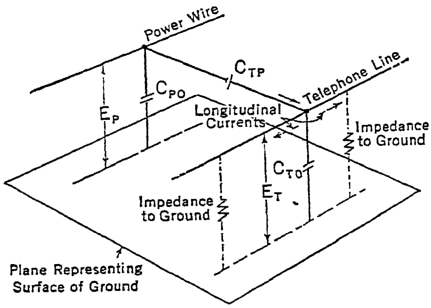


FIG. 5. Diagram Showing Couplings between Power and Telephone Wires for Electric Induction

$$E_T = \frac{C_{TP}}{C_{T0} + C_{TP}} E_P \qquad (1)$$

Where there are numerous power and telephone wires, capacitances exist between every possible combination of wires, and of wires and ground, resulting in a complicated network, but the principles involved are the same as in the simple case discussed above.

The potential of the telephone wire tends to be the same all along its length. If the wire is perfectly insulated from ground, extends only through the length of the exposure, and has no equipment on it, this potential is independent of the length of the exposure (the condition shown in Fig. 5 if the impedances to ground are neglected). This is true because, whereas all the capacitances in the above equation are proportional to exposure length, the ratio $C_{TP}/(C_{T0} + C_{TP})$ is independent of length. However, in practice, the circuits usually extend beyond the exposure and have equipment connected between them and ground, so that there are impedances to ground outside the exposure through which longitudinal current will flow. Since the impedance of C_{TP} controls the total longitudinal current, this current will be practically independent of the telephone-circuit impedances to ground and will be proportional to exposure length. It will also be proportional to the frequency of the harmonics in the inducing voltage. Hence, for given telephone-circuit impedance conditions (outside the exposure), the voltage to ground will be proportional to exposure length and to the frequency of the inducing voltage in a uniform and electrically short exposure.

In *magnetic induction*, the current in the power wire sets up a magnetic field which alternates at the frequency of the current. If a communication wire is located in this field, a *voltage is induced along it* which is proportional to the rate of change of the magnetic flux. This phenomenon is illustrated in Fig. 6. The voltage between the telephone circuit and ground varies from point to point along the circuit and depends on the distribution of the impedances to ground as well as on the distribution of the induced voltage. Also, since the voltage acts along the circuit and the part induced in each short length adds directly to those in all other short lengths, the total induced voltage is directly proportional to the exposure length in a uniform and electrically short exposure. Further, as the rate of change of magnetic flux is proportional to frequency, the induced voltage will be proportional to the frequency of the inducing current.

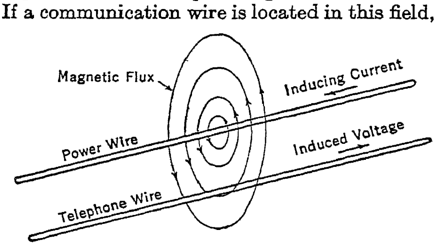


FIG. 6. Diagram Showing Coupling between Power and Telephone Wires for Magnetic Induction

In the foregoing, the factors discussed apply to both noise and low-frequency induction. However, these two general types of problems are discussed separately.

31. NOISE FREQUENCY INDUCTION

Noise frequency induction, in communication circuits, for a given situation of proximity, depends upon:

(*a*) Inductive influence of the power circuits, as determined by their electrical characteristics.

(*b*) Inductive coupling between the power and communication circuits, as determined by their physical relation.

(*c*) Inductive susceptiveness of the communication circuits, as determined by their electrical characteristics.

INDUCTIVE INFLUENCE. Two characteristics of a power system of primary importance in determining its inductive influence are wave shape and balance.

The wave shape of the voltage or current on a power line is a function of the magnitudes and frequencies of the harmonics, which may induce voltages of frequencies within the range ordinarily used in telephone circuits. Induced voltages at such frequencies have much greater interfering effects than the voltage induced at the fundamental frequency.

The approximate relative interfering effects of telephone line voltages and currents is shown in Fig. 7 for certain subscriber sets and instruments employed in the Bell System.

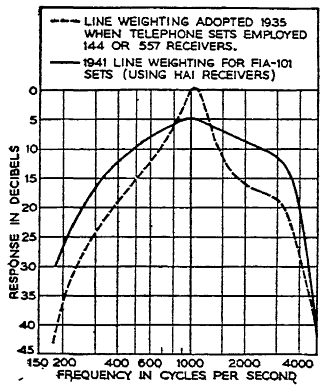
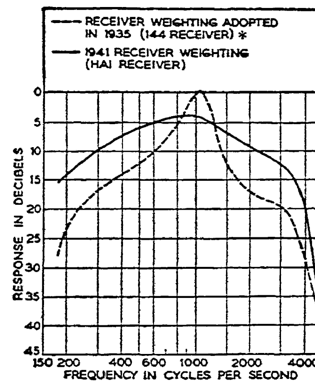


FIG. 7. Line Weightings for Telephone Message Circuit Noise (Relative Interfering Effects of Telephone Line Voltages or Currents) (Report 45, J.C.P.C.)



* Also applicable to 557 receiver for estimating receiver noise.

FIG. 8. Receiver Weightings for Telephone Message Circuit Noise (Relative Interfering Effects of Telephone Line Voltages or Currents at Receiver) (Reports 32 and 45, J.C.P.C.)

The solid curve shows the line weighting adopted in 1941 for the present-day anti-sidetone set having a relatively flat response over the useful voice-frequency range. This curve also takes into account (1) the possible utilization, in future telephone receiving elements, of a larger part of the wide-band toll transmission attained in the later-type carrier toll circuits, (2) a more uniform response within the transmission band, and (3) the attenuation at frequencies near 3000 cycles and higher, relative to that at 1000 cycles, caused by various types of switching trunks used in toll connections, as compared to the distortionless trunks used in the tests on which this curve is based. The two curves cannot be compared on an absolute basis, since the acoustic output per volt input at 1000 cycles is not the same for the different sets.

Receiver weighting curves for telephone message circuit noise are shown in Fig. 8 for certain instruments employed in the Bell System. These curves differ from the line weighting curves, principally as a result of taking into account the voltage loss, relative to that at 1000 cycles, caused by the trunk, loop, and telephone set, between the toll board (at which point the line weightings apply) and the telephone set receiver.

Each harmonic voltage or current induced in a telephone circuit by a paralleling power circuit will individually react in the telephone circuit, over which a conversation may be in progress, and in such a manner that the listener hears the combined effect of all the harmonics, which is termed noise. Different frequencies in this noise have different interfering effects, depending on the characteristics of the telephone circuits, type of receiver, the human ear, and other factors. Relative interfering effects (Fig. 8) are called noise weightings and have been determined by extensive tests.

Since the effects of electric and magnetic couplings are directly proportional to frequency, the relative noise influence of power-system voltages and currents is proportional to the product of noise weighting and frequency. For any frequency, this product times a constant gives the telephone influence factor, TIF. TIF weightings versus frequency are shown in Fig. 9 for three different periods, reflecting changes in transmission-frequency characteristics of telephone circuits and instruments from 1918 to 1941. The 1941 curve has not yet been standardized. It will be noted from Fig. 9 that the TIF values at 60 cycles are very low.

While a great deal of inductive coordination work makes use of single harmonic frequency data, there are cases where it is desirable to evaluate the overall influence of a voltage or current in a power circuit in terms of a single measure. Such a measure may be obtained by multiplying the magnitude of each harmonic present (amperes for current

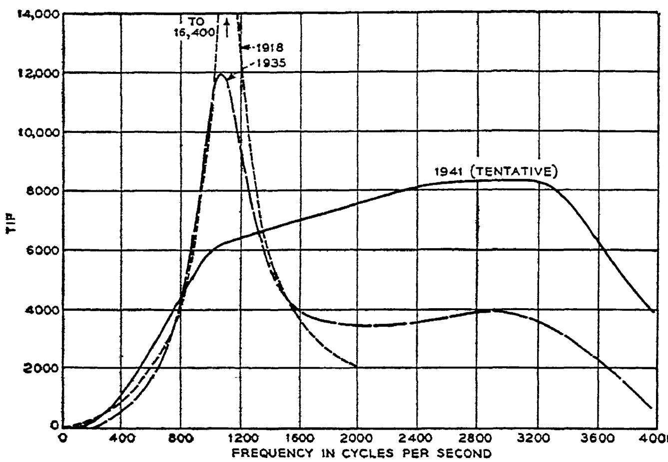


Fig. 9. TIF Weightings for Periods 1918, 1935, and 1941 (Courtesy A.I.E.E.)

and kilovolts for voltage) by its TIF weighting, and taking the rss value (square root of the sum of the squares) of these products. The result for current is the I·T product, and for voltage it is the KV·T product.

These products can be measured directly with a suitable noise-measuring set and a current or voltage TIF coupler, the latter having a transmission-frequency characteristic directly proportional to frequency.

The magnitudes and frequencies of the harmonic currents and voltages on a power line depend on the characteristics of the apparatus and associated equipment and on the impedance of the supply line at each of the harmonic frequencies. The wave shape at various points on the power system depends on the way in which the various harmonic currents and voltages are propagated over the system. Since power systems are usually very complex electrically, the propagation effects may vary greatly for different frequencies and for different systems.

It is impracticable to construct rotating electrical machinery or power transformers entirely free from harmonics, although marked progress has been made in this respect. Also, it is inherent in the operation of rectifying devices (and some other types of devices where the current is not directly proportional to voltage) that harmonics are produced. Generally speaking, the factors affecting the production of harmonics in these general classes of apparatus are as follows:

Motors and Generators. Harmonics are affected by the distribution of air-gap flux, variations in the air-gap flux due to the slots in the rotors and stators, the distribution of the windings on the armature, and, in multiphase machines, connections of the windings.

Transformers. The degree of saturation of the iron in the core affects the harmonics materially. In polyphase transformer banks, the connection of the transformers in the bank affects some of the harmonics, particularly the triple. In a 3-phase transformer the arrangement of the core also affects the triple harmonics.

Rectifiers. Owing to the flat-top wave shape of rectifier anode currents, the a-c line current, taken by the rectifier, has a step-type wave shape, as shown in

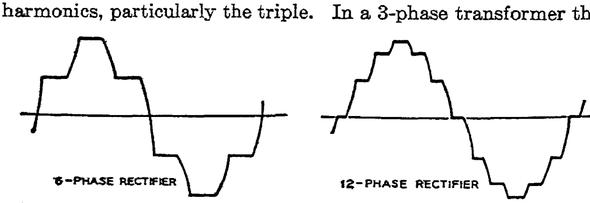


Fig. 10. A-c Line Currents Taken by 6 and 12 Phase Rectifiers (Courtesy A.I.E.E.)

Fig. 10 for a 6- and 12-phase rectifier. This wave shape results in the production of harmonic currents and voltages in the a-c line. Table 1 shows the order of these harmonics (multiples of the fundamental frequency) for various multiphase rectifiers with balanced operation, and also the reduction, with increase in phases, of the number of harmonics which are of importance.

The magnitude of any harmonic which is present with any particular number of phases is the same as for a 6-phase rectifier, assuming all other conditions to be the same. Thus, the magnitude of the twenty-third harmonic is the same for the 6-, 12-, and 24-phase rectifiers. Also, the magnitudes of the harmonic currents bear a definite relation to the rms value of the total rectifier current and decrease in value with increase in frequency. For a given kilovolt-ampere input to a rectifier, the higher the a-c line voltage the lower is the a-c line current, with a corresponding reduction in the magnitudes of the harmonic currents and usually in the harmonic voltages resulting from them. Furthermore, there are generally, under modern practices, fewer long closely coupled exposures with the high-voltage transmission circuits than with the lower-voltage power distribution networks.

Table 1. Harmonics Arising in Rectifiers

Orders of Harmonics in Line with Balanced Operation						Corresponding Harmonic Frequencies on 60-cycle System
Rectifier Phases						
6	12	18	24	36	48	
5						300
7						420
11	11					660
13	13					780
17		17				1020
19		19				1140
23	23		23			1380
25	25		25			1500
29						1740
31						1860
35	35	35		35		2100
37	37	37		37		2220
41						2460
43						2580
47	47		47		47	2820
49	49		49		49	2940
53		53				3180
55		55				3300
59	59					3540
61	61					3660

Note: Higher harmonics are also present for all types listed.

Phase control, employed on rectifiers for reducing the d-c output voltage below that obtained without phase control, is accomplished by retarding the firing point of the anodes in the alternating-voltage cycle, through grid or firing control. The power factor at the a-c line terminal of the rectifier transformer is lowered. The magnitude of the harmonic components increases in the a-c line current for a given kilowatt output of the rectifier. Also, with phase control, the anode currents have a steeper wave front at the beginning and end of the anode firing period, during commutation between successive anodes, resulting in harmonics of higher magnitude. Phase control should, therefore, be limited to actual requirements, particularly at full load and overload ratings of rectifiers, in order to limit the possible inductive interference.

Balanced and Unbalanced Currents. In a multiphase, *balanced* power circuit the voltages between the several phase conductors and between the phase conductors and ground, and also the several line currents, are vectorially equal to zero.

When the currents or voltages do not vectorially equal zero, they contain a set of single-phase components, all in the same phase relation, which are termed residual components. Any system of voltages or currents can be resolved into its balanced and residual components, and the effects of each can be analyzed separately. The balanced components are confined wholly to the phase conductors; the residual components act in a path consisting of the phase conductors and an external return as, for instance, a metallic neutral or through the earth. Since the coupling for the residual components is usually much larger than for the balanced components, the former are usually of greater importance in coordination problems.

Single-phase branches on three-phase distribution systems are, of themselves, inherently unbalanced. On grounded neutral systems the residual voltage on a single-phase branch is practically equal to the phase-to-neutral voltage. On isolated neutral systems the residual voltage on a single-phase branch depends on the particular system layout. The single-phase branches also introduce residual currents and voltages on the 3-phase system.

With the present methods of analyzing noise induction problems, the balanced and residual currents and voltages are usually considered separately. In the general case of exposures of overhead lines of the multigrounded neutral type to subscribers' cable circuits, a knowledge of the residual currents is sufficient, the effect of the balanced currents being relatively unimportant.

INDUCTIVE COUPLING. The coupling between power and communication circuits is determined by the degree of their proximity, but it may be greatly modified by the balance of the two classes of circuits to each other and by the proximity of grounded linear circuits or metallic objects.

Table 2

Types of Induction	Transpositions Tending to Reduce Induction {T = Telephone, P = Power}
A. Metallic-circuit (direct)	
1. From balanced currents	T
2. From balanced voltages	T
3. From residual currents	T
4. From residual voltages	T
B. Longitudinal-circuit (indirect metallic-circuit) *	
5. From balanced currents	P
6. From balanced voltages	P
7. From residual currents	P } Only if residuals are
8. From residual voltages	P } thereby reduced.

* This component of induction can result in noise in the metallic circuit, because of the reaction of such longitudinal induction upon self-unbalances (high-resistance joints or leakage) or mutual inductance or capacitance unbalances to other wires on the line or to ground.

In determining coupling, it is desirable to differentiate between the effects of the balanced and residual components in the power circuit, between the effects of voltages and those of currents, and, on the telephone line, between induced voltage which acts directly in the metallic circuit, termed *metallic-circuit induction*, and that which acts in the circuit composed of the wires in parallel with ground return, termed *longitudinal-circuit-induction.*

Types of Induction. Eight components of power induction, shown in Table 2, need to be considered in noise induction problems. These components vary in their importance, as noise factors, for different situations and for the reasons discussed above.

Transpositions are employed in open wires within exposures (parallels) between power and communication circuits as an aid in neutralizing induced power-circuit noise in the latter circuits. Transpositions are required generally in open-wire communication circuits, whether or not power-circuit exposures exist, to limit intercircuit cross-talk.

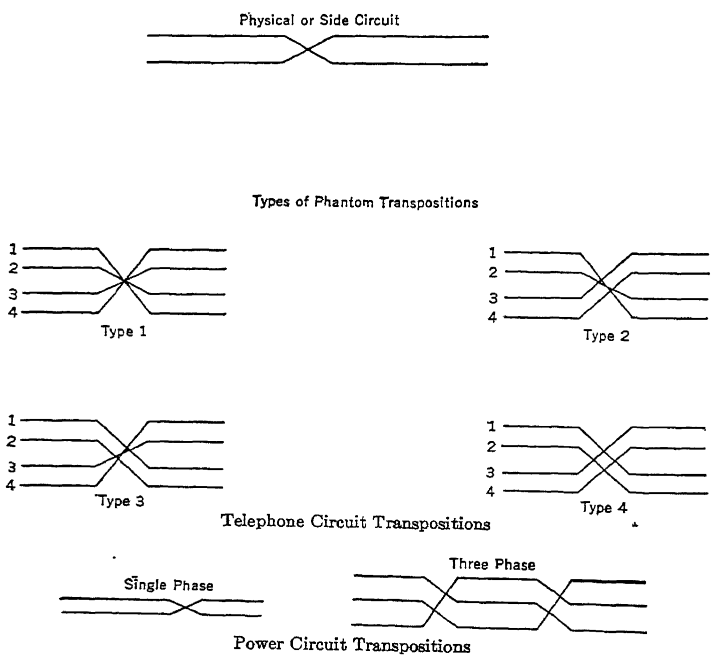


FIG. 11. Diagrams Showing Wire Arrangements at Transposition Points

Transpositions are made by interchanging in a uniform manner the positions of the wires comprising a circuit, so that each wire of the circuit occupies all the pin positions occupied by the circuit, for distances (usually equal) as determined by the transposition design. Figure 11 shows the changes employed in the position of the wires (both telephone and power) at transposition points. Telephone transpositions may be of the physical or side circuit types involving two wires or of any one of four types of phantom transpositions involving four wires. Power transpositions may be of the single-phase type involving two wires or of the 3-phase type, involving three wires.

Identical 3-phase power transpositions, when placed at the $^1/_3$ and $^2/_3$ points in a given uniform section of power line, establish a *power circuit barrel,* since each wire is rotated 120° in phase position and in the same direction of rotation at each transposition. Single-phase power transpositions rotate the wires 180° in phase position. Likewise, for metallic-circuit induction, telephone transpositions change the phase of the induction by 180°.

Figure 12 shows a simple arrangement of telephone and power transpositions within a unit exposure length, L, containing one power barrel. In this arrangement, which is

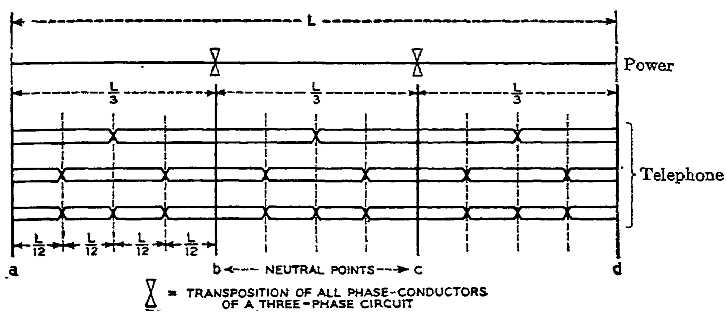


FIG. 12. Simple Balanced Arrangement of Telephone and Power Transpositions in a Unit Exposure Length, L (Courtesy Bell System)

commonly employed, the 3-phase power transpositions are located at neutral points with respect to the telephone transpositions, and the telephone transpositions are so arranged as to reduce the direct metallic-circuit induction (from both balanced and residual components) within each section of exposure between power transpositions, as well as to limit inter-cross-induction.

Telephone circuit transpositions tend to

(*a*) Reduce intercircuit mutual effects, known as cross-induction or cross-talk.

(*b*) Reduce direct metallic-circuit induction from both balanced and residual components of power circuits, within exposures, especially when coordinated.

(*c*) Balance the two sides of the telephone circuit with respect to earth and with respect to all other wires on the telephone line, considered as one longitudinal circuit (Sigma).

These transpositions, by themselves, are not effective in reducing the longitudinal-circuit induction from either the balanced or residual components of the power-circuit voltages and currents. However, within an inductive exposure, such transpositions tend in some measure to equalize these inductive effects by exposing each conductor of the telephone circuit equally to the power-circuit influences.

For telephone-circuit transpositions to be reasonably effective in reducing metallic-circuit induction, the relation between the power and telephone circuits within each coordinated section of exposure (with respect to each other, to ground, and to other circuits present) must be substantially uniform. Thus, points of discontinuity within an exposure, such as sharp changes in separation, crossings, or changes in power-circuit configuration, must be considered in inductive coordination work.

Power-circuit transpositions, within the exposure, tend to reduce the longitudinal-circuit induction from balanced voltages and currents and, as a result, that component of metallic-circuit noise arising mainly outside the exposure due to the action of the longitudinal-circuit induction (from balanced components) upon any unbalances affecting the telephone circuits.

Notes. 1. Power transpositions, in the usual case, do not appreciably affect the direct metallic-circuit induction.

2. That part of the direct metallic-circuit induction which results from residual voltages and currents is not affected by power-circuit transpositions except in so far as such transpositions may reduce the power-circuit residuals

INDUCTIVE SUSCEPTIVENESS. The degree to which telephone transmission is adversely affected by noise-frequency induction depends not only upon the magnitudes of the induced noise voltages, as determined by influence and coupling factors, but also upon the susceptiveness factors of the telephone system. These include the manner in which the induced voltages and currents are propagated to the circuit terminals, together with the reactions of the circuit unbalances (thus relating the current in the terminal apparatus to the induced voltages), the sensitivity of the receiving apparatus, and the operating power level of the telephone circuits.

Propagation Effects and Balance. Important differences exist with respect to propagation effects and balance between open-wire and cable circuits and between toll and exchange systems.

As pointed out in the discussion of coupling, only the magnetically induced longitudinal voltages and currents are important, under the conditions usually encountered, in producing noise in telephone cable circuits. Because of the negligible effects of electric induction and direct metallic-circuit induction and because of the important shielding effects exerted by the cable sheath and the various telephone circuits on each other, telephone cable circuits are much less susceptive than open-wire circuits.

In open-wire telephone systems, consideration must be given both to electric and magnetic induction and to voltages induced directly in the metallic circuit as well as to those induced in the longitudinal circuit. In a line composed of a number of circuits, the currents set up in any one circuit depend not only upon the voltage induced in that circuit and its impedance but also upon the currents and voltages which are set up in the other telephone circuits on the line. It is not possible, therefore, to calculate precisely the induced currents merely from a knowledge of the magnitude of the currents and voltages on the power circuits and the coupling between the power circuits and isolated pairs of wires on the telephone line, considered independently.

Estimates of receiver noise for a particular type of subscriber set and receiver connected to a given line circuit, in which noise currents are assumed to be present, may be made as discussed below.

On the basis of the definition of reference noise, in absolute terms, as an electrical power of 10^{-12} watt at a frequency of 1000 cycles dissipated in a receiver (for all types of receivers), Table 3 gives the currents in three types of receivers and the voltages across them at 1000 cycles for reference noise (0 db) and representative values of the receiver impedances.

Table 3

Receivers (W.E. Co. types)	Current, micro-amperes	Voltage, micro-volts	Impedance, ohms
144	0.0795	21.60	$158 + j221 = 272/54.5°$
557	0.1414	19.70	$50 + j130 = 139/69°$
HA1	0.1210	16.56	$68 + j118 = 136/60°$

At frequencies other than 1000 cycles, the voltage for reference noise is the voltage across the receiver producing an interfering effect equal to that produced by the reference noise voltage at 1000 cycles. For a given receiver, the voltage for reference noise at a frequency other than 1000 cycles may be obtained from the 1000-cycle reference voltage and the proper noise weighting curve for receiver voltages (if available). Figure 8 shows such weighting curves for the three receivers listed in Table 3. Also, for a given receiver, the receiver current for reference noise at a frequency other than 1000 cycles may be obtained from the relation between the voltage for reference noise across the receiver at the given frequency and the receiver impedance at the same frequency. Values of receiver currents and corresponding voltages for reference noise at various odd harmonics of 60 cycles, obtained in accordance with the above procedure, are given in Fig. 13.

From the current for reference noise at a given frequency f and the susceptiveness factor at the same frequency for a subscriber set, the receiver noise (in decibels) for 1 volt to ground on the set at that frequency is given by the expression

$$N_f \text{ (receiver noise in db)} = 20 \log_{10} \left(\frac{I_f}{I_r}\right)$$

where I_f = receiver current in microamperes per volt to ground at the subscriber set at frequency f = susceptiveness factor for set involved.

I_r = receiver current in microamperes (for the type of receiver involved) for reference noise at frequency f (Fig. 13 for three different receivers).

Values of N_f for the 144 receiver, calculated by the above expression, require no corrections, since present receiver noise transmission impairments are related to noise magnitudes in this type of receiver. Values of N_f so calculated for the 557 and HA1 types of receivers are not representative of the interfering effect on the present standard decibel scale of noise for the 144 receiver. For the final results of all three receivers to be on approximately the same basis as to noise transmission impairments, it is necessary to adjust the N_f values of the 557 and HA1 receivers by factors of +4db and −4db, respectively, which factors are based on the results of judgment tests. By such adjustments, the readings are in *dba*, a term that can be interpreted as representing the quantity that results after properly "adjusting" the decibel reading of the noise measuring set. When the reading is thus adjusted all noise results will be on a common basis.

Receiver →	Currents			Voltages			Receiver →	Currents			Voltages		
	144	557	HA1	144	557	HA1		144	557	HA1	144	557	HA1
Frequency, cps	Microamperes			Microvolts			Frequency, cps	Microamperes			Microvolts		
180	4.21	9.29	1.41	530.	483.	60.8	1620	0.304	0.505	0.147	112.	101.5	26.2
300	0.954	2.11	0.566	144.	131.5	31.7	1740	.334	.549	.154	128.	116.9	28.2
420	.586	1.27	.338	106.	96.6	23.6	1860	.365	.586	.157	146.	133.	29.9
540	.40	0.82	.240	80.1	73.	20.1	1980	.373	.604	.158	159.	145.	31.5
660	.281	.542	.181	60.6	55.3	18.1	2100	.376	.598	.156	167.	152.	32.9
780	.180	.34	.149	43.3	39.5	17.1	2220	.377	.593	.154	167.	159.	34.4
900	.109	.205	.131	28.7	26.2	16.7	2340	.377	.591	.152	181.	165.	35.8
1020	.077	.138	.120	21.2	19.5	16.7	2460	.377	.588	.150	185.	172.	37.2
1140	.094	.161	.123	27.1	24.7	18.5	2580	.377	.580	.148	193.	178.	38.9
1260	.138	.235	.129	42.4	38.6	20.5	2700	.377	.569	.155	200.	182.	40.8
1380	.2	.343	.134	66.2	60.3	22.3							
1500	.258	.445	.140	91.2	83.6	24.3							

FIG. 13. Receiver Currents and Voltages Corresponding to Reference Noise.* (Report 46, J.C.P.C.)

* Weighted currents and voltages based on representative receiver impedances and definition of reference noise as 10^{-12} watt dissipated in each receiver at 1000 cycles.

Power Level and Sensitivity. The susceptiveness of telephone circuits to induced noise from power-supply circuits or other outside influences depends to a considerable degree on the levels of the voltages and currents used in speech transmission and in the efficiency of the telephone terminal apparatus in converting electrical into sound power. Power level, as discussed here, refers to the level of speech currents (with respect to reference level of speech transmission) and not to a level measured in watts.

Speech power, and consequently electrical power generated by a subscriber set telephone transmitter, which is actuated by speech power, varies over a wide range of values and frequencies. This variation will occur with any one speaker and is usually different for different speakers. Electrical power also varies in different toll and exchange circuits, owing to the different types of lines and apparatus encountered and line length. In one investigation, the average acoustic power of speech produced by 16 talkers was of the order of 10 microwatts. The average ratio of the maximum instantaneous power to the average power, for the various vowel sounds only, was of the order of 15 to 1, whereas the ratio of the maximum to average power for a continuous sine wave is 2 to 1, thus showing the much wider variation of speech power as compared to generated electric power.

The power on commercial telephone circuits is conveniently determined by a device known as the volume indicator. This device consists primarily of a rectifier-type indicating meter of specific dynamic characteristics. When measuring speech power, the meter deflections fluctuate continually in response to the variations of speech power. Because of this varying deflection, it is necessary to specify a standard method of interpreting the indications, which involves adjusting a calibrated potentiometer, associated with the meter, to maintain the meter needle deflections approximately in a specified range on the scale. When these deflections correspond to the specified reading, with the volume indicator connected across 600 ohms, and the potentiometer is set at zero, the power indicated by the meter in the circuit being measured is zero VU (volume units).

Since, by the action of the carbon granule transmitter, a relatively large amount of electrical power is controlled by movement of the diaphragm, the electrical power delivered to a telephone circuit is much greater (of the order of several hundred times) than the acoustic power delivered to the transmitter. This amplification is of value in maintaining speech power at satisfactory levels above induced power current levels.

Transmitter developments have tended to raise the response level for some frequencies in order to give a more nearly flat frequency characteristic over the voice range without materially increasing the maximum level output of the transmitter. Such a characteristic improves the articulate qualities of the speech and hence effective transmission.

Receiver developments have also increased receiver efficiency over the earlier periods of operation, in addition to flattening its response over the voice range. However, its actual efficiency (electrical power input to acoustic power output) is relatively low. The power loss in the receiver at low frequencies, such as 25 and 60 cycles and their lower harmonics, is much greater than at frequencies in the higher voice range. The combination of amplification of voice currents in the telephone transmitter with noise current loss in the receiver permits delivering to the subscriber satisfactory speech power while keeping within tolerable limits the sound levels due to ordinary amounts of induced noise.

Speech levels on toll circuits are, in general, maintained at about specified levels by means of telephone repeaters (not usually employed on exchange circuits), which amplify the speech as well as any noise currents and voltages induced in the toll circuits. These repeaters are usually spaced at suitable intervals on a toll circuit to provide the proper speech level without overloading the amplifier and without causing cross-induction between adjacent circuits, due to excessive or inadequate levels.

The trend toward improved balance of party-line subscriber sets and some parts of central-office circuits permits increasing the speech-to-noise level ratio in exchange plant. Any betterments of this type, within limits, which increase the signal-to-noise ratio tend to decrease the noise effects in telephone circuits.

32. NOISE INDUCTION MITIGATION

Noise induction mitigation usually involves careful consideration of at least several of a large number of factors, which may be broadly classified under the headings: (*a*) influence factors; (*b*) coupling factors; (*c*) susceptiveness factors.

COOPERATIVE PLANNING in connection with the design and location of lines and systems is of great importance. These cooperative plans generally are directed toward: (*a*) coordinating the locations of lines; (*b*) incorporating in the design of both systems those features which will limit the influence and susceptiveness.

By cooperative planning, not only can the number of exposures be limited but also the general designs of the systems can be made such that treatment of individual exposures is materially simplified. Furthermore, in connection with new construction or changes in either system, coordination can be considered before expenditures or other commitments are made.

CONTROL OF INFLUENCE FACTORS. Residual currents and voltages of the triple-harmonic series can be controlled by one or more of the following means:

(*a*) Opening the neutral-to-ground connection of the machine or transformer bank where the triples originate. This can be done only where other system grounding arrangements, adequate from the standpoint of power system stability and relaying, are available.

(*b*) Opening the neutral-to-ground connection of transformer banks through which triples from another source complete their path. System stability and relaying must also be considered in this connection.

(*c*) Providing a path for triples (such as with a wye-delta bank), which tends to shunt them out of the exposure.

(*d*) The use of wave traps (anti-resonant circuits tuned to the important harmonics), reactances, or other devices in the neutral-to-ground connections at locations where triples originate.

(*e*) The use of transformer connections (such as wye-delta or delta-delta) through which triples will not pass.

(*f*) The use of rotating machines, which have a low influence factor.

Note: The last two measures are usually of the greatest importance in connection with cooperative advance planning, since to change existing equipment to these types may be unduly expensive.

Residual currents and voltages of the non-triple harmonic series can be controlled by:

(*a*) Reduction of the unbalance which gives rise to them, as by changing single-phase taps to 3-phase, balancing single-phase taps among the phases of the 3-phase line, or balancing loads among the 3 phases, where neutrals are multigrounded.

(*b*) Absorption by wye-delta banks or other means.

In addition to the design of power apparatus to limit harmonics as far as practicable, and the avoidance of excessive magnetic densities, frequency selective devices, to filter out harmonics, have been used in some situations, for example:

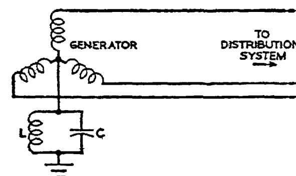
(*a*) On the d-c sides of trolley rectifiers.

(*b*) On the a-c sides of rectifiers.

(*c*) Across the terminals of rotating machines to reduce important balanced harmonic currents and voltages.

(*d*) In the neutral-to-ground connections of generators, synchronous converters, and other power devices, to reduce triple-harmonic voltages and currents, an example of which is shown in Fig. 14.

CONTROL OF COUPLING FACTORS. One highly satisfactory method for the control of inductive coupling, when it can be employed, is the complete physical separation of the power and telephone lines. However, in built-up communities both types of service are required by the public, making it necessary to utilize the same routes for distribution. For intercity toll lines, which are usually important backbone routes for long-haul communication traffic, frequently reasonable separation from power transmission systems can be obtained, particularly with the proper cooperative advance planning.



Frequency which trap is tuned to suppress

$$= F = \frac{10^3}{2\pi \sqrt{LC}}$$

L = inductance (henrys)
C = capacitance (microfarads)

FIG. 14. Wave Trap in Grounded Neutral of Line-connected Generator (Report 12, J.C.P.C.)

Where power line parallels with open-wire communication lines (toll or exchange) are created, transpositions are usually effective in controlling the resulting inductive couplings, and these may be required on a coordinated basis.

Whether the coordination of telephone transpositions with the discontinuities in the exposure, or the use of power-circuit transpositions, or both, is desirable in a specific case will depend on the relative importance of direct metallic-circuit induction and longitudinal-circuit induction acting on telephone-circuit unbalances, and on the importance of the induction from the balanced and residual components of the power-circuit voltages and currents.

Telephone transpositions must also be effective in controlling cross-induction (cross-talk) between telephone circuits. Standard transposition arrangements have been devised to meet this requirement for different classes of open-wire facilities. Two of such arrangements, which are available, are shown for four arms of wire in Figs. 15 and 16.

Sometimes, unavoidable irregularities occur in the spacing of poles, in distances between power and telephone circuits, in the presence of shielding objects such as other communication lines and trolley systems, and in heights of the circuits, which it is not practicable to take into account in the transposition design. Where these irregularities are large, the effectiveness of the transposition arrangements is correspondingly impaired.

Possible benefits are illustrated by noise measurements in the rural joint-use exposures, established in Alabama, Minnesota, and South Dakota (see article 29). The ratio of receiver noise to noise-to-ground is a good indicator of the effectiveness of noise-reduction measures, and such measurements, made under normal conditions, gave ratios of the order of -36 db (about 1 to 60 in voltage), where the IT (effective power harmonic currents in the line times the voltage interference factor) is fairly high. A ratio of this order is adequate with respect to noise, except in extreme exposures, and should result in noise on rural joint-use circuits comparable to that on urban party-line circuits in cable.

The favorable noise results in these instances are due partly to the effectiveness of the transposition scheme used, which employs frequent, point-type transpositions on tandem brackets and an average wire spacing of only about 7 in. Also in many of the exposures the power circuit is of the single-phase common-neutral type with vertical configuration, for which, in joint use, the direct metallic induction into a horizontal telephone circuit is inherently low. Of course, low values of receiver noise are not obtainable unless the telephone circuits are well balanced.

Shielding. When either or both of the telephone and power facilities are enclosed in metallic cable sheath, having a relatively low resistance to ground, the sheath acts as an effective shield against both electric and magnetic induction. For magnetic induction, the induced longitudinal currents flow along the sheath, which has a finite resistance per unit of length, different from the enclosed conductors, and complete shielding from such currents is thus not obtainable.

Present trends are toward bonding local distribution aerial cable to multi-grounded power neutrals, where these neutrals are well bonded to extensive public water systems, since, in the usual case, the cable sheath becomes quite closely associated with the common power neutral through the telephone drops, station protector grounds, power secondary services, and telephone company practices of placing protector blocks between working lines and the cable sheath. Frequent bonds between the common neutral and cable sheath (about every 1/4 mile or less) are not only advantageous from a protection standpoint but also useful for noise mitigation.

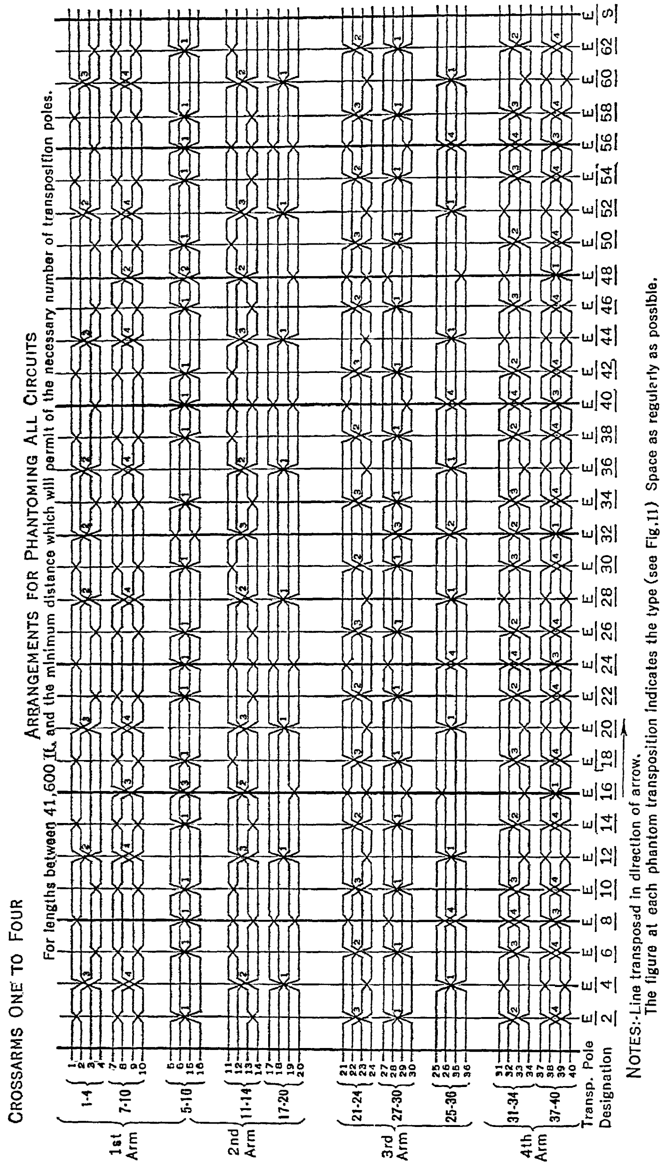
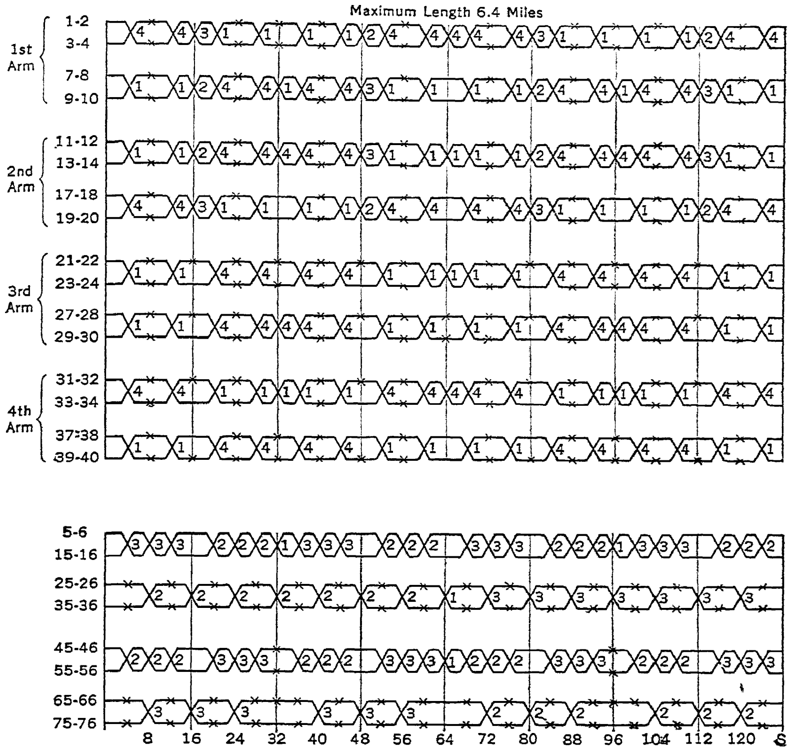


Fig. 15. Typical Transposition Scheme for Phantomed Circuits Suitable for Use in Inductive Exposures (Courtesy Bell System)

Such bonding usually results in an average reduction of subscriber receiver noise of the order of 2 to 1 for station sets equipped with single condensers and low-impedance ringers to ground. The average reduction in noise to ground is about 3 to 1. In non-public water-pipe areas the noise reductions obtained by bonding to the neutral depend upon the resistance to ground of the neutral conductor: the lower the resistance, the greater the reduction.

For power systems of the 2.3/4.0-kv type, increased potentials on the telephone sheath may generally result from bonding to the power neutral, but tests indicate no cable insula-



NOTE:—The figure at each phantom transposition indicates the type (see Fig. 11).

FIG. 16. Typical Transposition Scheme (Improved Type) for Phantomed Circuits Suitable for Use in Inductive Exposures (Courtesy Bell System)

tion failures due to this potential increase. In trolley areas, additional direct current will usually be transferred to the underground telephone cable sheaths because of power neutral bonding, but the increase in sheath current will probably not require additional corrective measures for electrolysis.

The effectiveness of shielding is expressed in terms of a *shield factor*, which is the ratio of the noise in the shielded to the noise in the non-shielded condition. Figure 17 shows observed shield factors in four public water system areas obtained with telephone cable sheaths bonded to a multigrounded power circuit neutral at a number of places.

CONTROL OF SUSCEPTIVENESS. In toll circuits, which are designed to be symmetrical with respect to earth, the reduction of unbalances is usually a matter of correcting conditions which are the result of deterioration or maintenance, although situations occasionally arise where the design of apparatus is involved. The former includes:

(*a*) High-resistance joints. The remedy is to make a new joint.

(*b*) Leakage caused in open-wire circuits

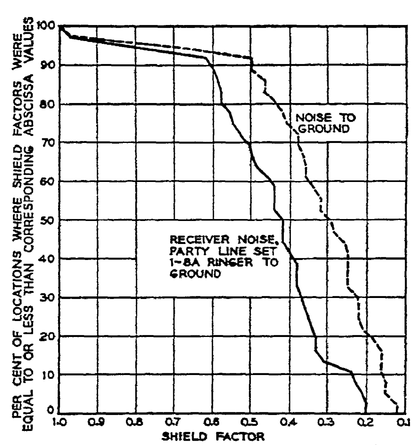


FIG. 17. Shield Factors—Telephone Cable Sheath Bonded to Power Circuit Neutral (Courtesy Bell System)

by trees or broken or missing insulators. In cable circuits, leakage is usually caused by moisture entering at a sheath break.

(*c*) Capacitance unbalances in open-wire circuits. The remedy involves a careful check of the transpositions either by inspection or by suitable electrical testing means.

(*d*) Incorrect connections or unbalanced arrangements of apparatus at terminals. For example, composite sets should not be placed on one side circuit of a phantom group without similar sets being placed on the other.

(*e*) Incorrect connections in entrance cables, such as split pairs or quads.

In regard to design, the causes of unbalances are usually in terminal apparatus, where the elements in the two wires of a circuit (or the two sides of a phantom group) are not sufficiently alike in impedance at noise frequencies. Modern apparatus is usually designed so that the series impedances and admittances to ground are very closely similar for the wires in a pair or quad. In some of the older designs, however, the degree of balance may not be sufficiently high. Sometimes, improvements can be secured by selecting among existing equipment the units having similar characteristics and grouping them together on pairs or quads. In some instances, there may be unbalances in entrance cables or office cabling, for example where phantomed circuits are routed through non-quadded cables.

In the exchange plant, unbalances due to connections of ringers to ground can be reduced by the use of high-impedance ringers or by other subset apparatus with improved balance. Central-office-circuit unbalances may need to be improved by modifications in or replacements of existing apparatus of the older types and of unsymmetrical design. The more recently designed central-office equipment is better balanced, and further improvements in this respect may be expected.

Sometimes improvement can be secured by inserting a balancing impedance in the other side of the circuit. Through cooperative advance planning, apparatus having improved balance can be introduced in an orderly manner.

Isolation of equipment unbalances can sometimes be secured by inserting between the apparatus and the line a well-balanced repeating coil without ground connections or with ground connections so arranged that longitudinal voltages and currents are not transmitted. It is necessary to arrange the circuit so that signaling and supervision will not be interfered with. A less effective but sometimes adequate method of isolation consists of inserting between the apparatus and the line a well-balanced coil so connected as to be non-inductive to the metallic circuit but to present a high longitudinal impedance. A well-balanced repeating coil, with the windings suitably connected, will frequently serve this purpose. This method has the advantage that it can be readily arranged so as not to interfere with d-c signaling and supervision. In both the toll and exchange plants, it is frequently necessary to guard against interconnection of balanced and unbalanced circuits through cord circuits not containing repeating coils, since such a connection would be unbalanced. This can be done by avoiding the use of such cord circuits for these connections or by isolating the unbalanced lines by repeating coils.

Figure 18 shows a schematic of a typical local step-by-step connector circuit in the talking condition, with a long line circuit inserted adjacent to the line in order to prevent longitudinal current flow from the calling line causing noise, due to possible unbalances in relays *A* and *D* and in the 2-μf series condensers. This is one means of preventing longitudinal circuit noise reaching the called subscriber from a connected line which is affected by power induction.

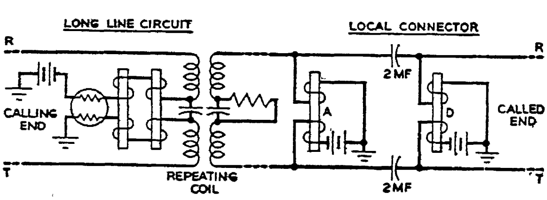


Fig. 18. Schematic of Typical Local Step-by-step Connector Circuit in Talking Condition Associated with a Long Line Circuit (Courtesy Bell System)

Transmission Levels. The effect of noise is lowered as the power level of the voice currents of telephone circuits is raised. In the toll plant, this fact has had a marked bearing on the sizes of wire used and the location of repeater stations. One of the limitations on the degree to which levels can be raised on toll circuits by repeaters is the difficulty of avoiding cross-talk between circuits on which there are large level differences. Subject to this limitation, however, advantage may sometimes be taken in specific situations of allocating repeater gains in such a way as to use the highest practicable level through inductive exposures.

In the exchange plant, telephone repeaters are normally not employed, so that the control of levels in connection with specific noise situations is a less practicable procedure.

However, the desirability of utilizing the highest practicable levels has had an important bearing in the development of instruments, cables, and other facilities.

Other Devices. In special cases, neutralizing transformers, resonant shunts, or resonant drainage to ground, applied to the telephone circuits, offer possibilities as coordinative measures for the reduction of noise induction.

The neutralizing transformer is employed, primarily, in local communication circuits serving power stations to limit voltages to ground at such stations when power-line faults

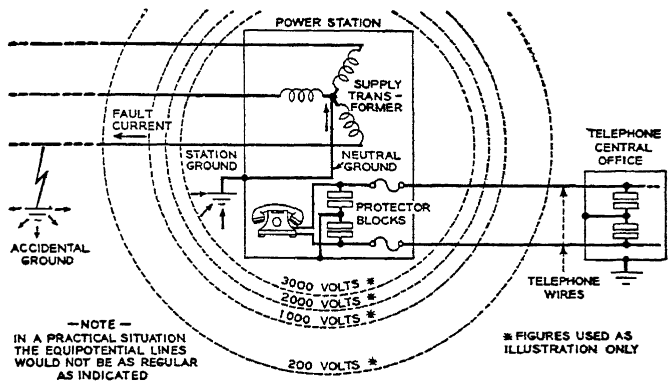


FIG. 19. Diagram Illustrating Rise in Station Ground Potential Due to Power Line Fault (Report 44, J.C.P.C.)

occur. Without the transformer, the rise in potential of the power station ground, and consequently the telephone set protector ground, would frequently be enough to break down the protector blocks and disable telephone service, as shown in Fig. 19. With the transformer, differences of potential between the communication circuit and nearby grounded structures are materially reduced, as indicated in Fig. 20, thus minimizing hazard to personnel, cable troubles, and service interruptions.

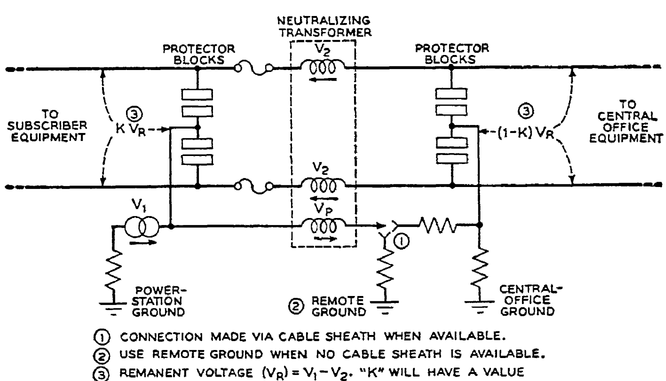


FIG. 20. Diagram Illustrating Action of Neutralizing Transformer in Neutralizing Voltages on a Circuit Subject to Ground Potential Rise (Report 44, J.C.P.C.)

Ground Return Circuits. Ground return telephone circuits, being inherently unbalanced to ground, require methods of coordination with power circuits different from those employed for metallic telephone circuits, such as previously discussed in this section. Ground return power circuits are not considered standard practice but are frequently operated as single-phase grounded neutral circuits in rural power systems. Coordinative measures are available for this type of power circuit with metallic telephone circuits, such as control of power-circuit influence factors and transpositions, drainage coils, or high impedances (to predominating harmonics) inserted in the telephone circuit.

33. LOW-FREQUENCY INDUCTION

Low-frequency induction, under *normal balanced power system operation,* rarely creates a noise problem in communication systems. However, under some normal operating conditions, the induced ground-wire (for lightning protection) current flow at the power-circuit fundamental frequency may be objectionable, particularly in telegraph operation.

When an abnormal condition, such as a grounded phase wire, occurs on a power circuit, relatively large currents at the fundamental frequency flow from the power circuit to ground for *grounded* power systems. This may result in excessive low-frequency induction in any paralleling communication circuits. This condition also obtains in an *isolated* power system when two or more phases are faulted.

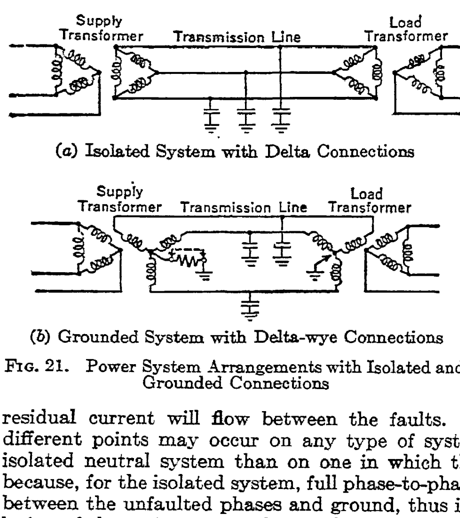
The magnitude of the voltage induced in a telephone circuit at the time of a power-circuit fault depends chiefly on the magnitude of residual currents and on the exposure conditions.

Residual Currents. If 1 phase of a 3-phase power line develops a fault to ground the currents in the 3 phases become unequal, and their vector sum, which is the residual current, is no longer zero. In most low-frequency induction problems, residual current is far more important than residual voltage.

The relatively large inductive influence of residual current is due to the fact that it exists in a circuit consisting of the line conductors in parallel as one side and the earth as the other side. Since much of the return current is deep in the earth, its neutralizing action is small.

The chief factors that determine the magnitude of the residual currents are (1) the power-circuit voltage, (2) the line and apparatus impedances, (3) the fault and earth impedances, (4) the impedances of the neutral ground connections, (5) the type of ground wire, if used, and (6) the circuit configuration including ground wires.

In analyzing the impedances controlling fault current, two general types of power systems must be considered: the *grounded neutral*, and *isolated neutral* systems. These are illustrated in Fig. 21. In the *grounded neutral* system, the neutrals of one or more transformer banks are grounded directly or through impedance so that, in the event of a fault, a path for current is established from the fault through the earth and returning through the neutral-to-ground connections. In the *isolated neutral* system, there are normally no grounds on the system so that, in the event of a fault, the path for fault currents is through the capacitances of the unfaulted phases to earth. Hence, for a single fault, the fault current is limited to the charging and leakage current.



(*a*) Isolated System with Delta Connections

(*b*) Grounded System with Delta-wye Connections

FIG. 21. Power System Arrangements with Isolated and Grounded Connections

If, in an isolated neutral system, a second fault to ground occurs on another phase while the first persists, a residual current will flow between the faults. Simultaneous faults on two phases at different points may occur on any type of system but are more likely to occur on an isolated neutral system than on one in which the neutral is solidly grounded. This is because, for the isolated system, full phase-to-phase or possibly higher voltage is impressed between the unfaulted phases and ground, thus increasing the voltage stress on the insulation of the entire system during the time of fault.

A system grounded through a neutral impedance has some of the characteristics of an isolated neutral system. Generally, the addition of neutral impedance tends to reduce the fault current, this effect being proportionately larger for faults near the neutral grounding points. This reduces the voltage induced on nearby telephone lines and may have advantages to the power system. On the other hand, increasing the neutral impedance may introduce problems in power system relaying. It may also increase overvoltages on the power system.

The duration of residual current is also important, since the length of time that the induced voltage persists has important reactions on its effects. For example, with the carbon block protectors, the chance of their becoming permanently grounded, with consequent interruption of service, depends not only upon the amount of current through the protector but also upon its duration. Likewise, other effects which are described later are materially affected by the duration of the induced voltage. Since, except for self-clearing faults, the duration of fault current is determined by the time of operation of the power-current interrupting devices, their reliability and speed of operation are important.

COUPLING FACTORS. Coupling is proportional to the length of the exposure for uniform separation. It varies with separation in a manner which is affected, among other things, by the structure of the earth. This effect can be summarized as follows:

Under the conditions of low-frequency induction, the telephone wires comprise one side of a loop, the other side of which is the earth. Likewise, the power wires comprise one side of a loop, the other side being the earth. The magnetic coupling between two parallel loops at a given separation increases as the size of the loops increases. The sizes of the loops are determined by the distribution of the return current in the earth. Generally speaking, the greater the resistivity of the earth, the more the current will spread and the greater will be the coupling to an adjacent circuit.

The effect of earth resistivity on coupling is much greater for wide separations than where the lines are close together. Consequently, with high-resistivity earth, not only is the coupling higher at all separations than with low-resistivity earth, but (except for very wide separations) the percentage reduction secured by increasing the separation a given amount is smaller.

Figure 22 shows, for several earth resistivities, the variation of mutual resistance, reactance, and impedance with separation, based on calculations using Carson's formula and a frequency of 60 cycles, without shielding.

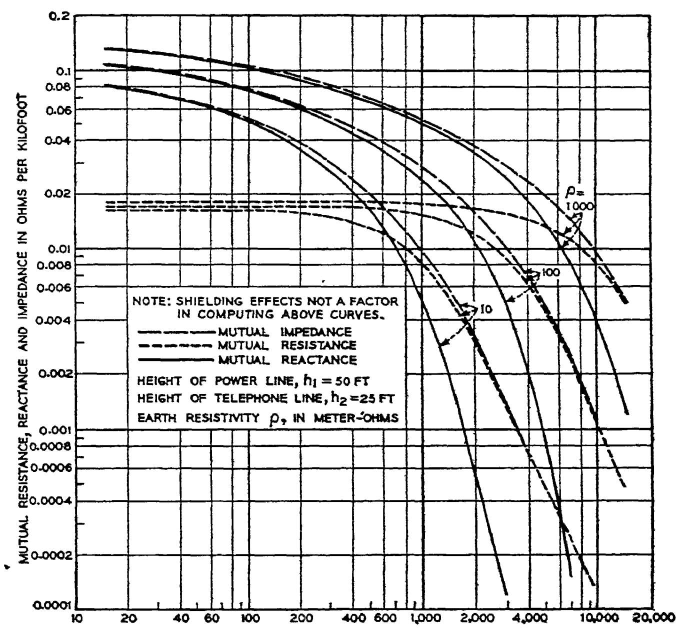


FIG. 22. Variation of Mutual Resistance, Reactance, and Impedance with Separation (at 60 cycles) (Report 14, J.C.P.C.)

SHORT-CIRCUIT CURRENTS, during power-line faults, may be calculated by several *approximate* short-cut methods, one of which may briefly be described as follows:

(*a*) Prepare a diagram of the power-system network, showing lengths, location, and kva capacities of large generating sources and transformer banks; location, kva capacities, and connections of grounded neutral transformer banks; and magnitudes of neutral impedances, if any.

(*b*) Show the location and separation of the telephone-circuit exposure or exposures to the power-system network.

(*c*) Show the line-to-line operating power voltage.

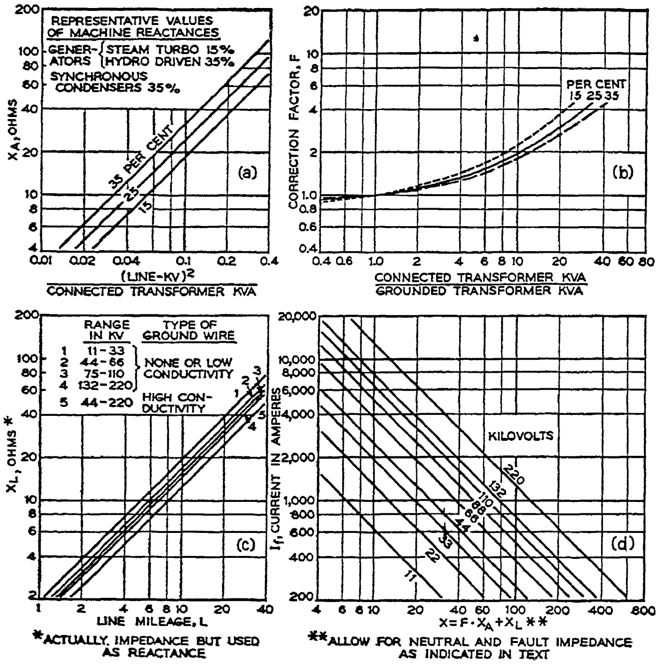


FIG. 23. Curves for a Quick Approximation of Fault Current Resulting from Single-phase Faults to Ground on Three-phase Systems (Courtesy Bell System)

(d) Using the various curves a, b, c, and d in Fig. 23, determine for a given situation:

1. X_A from $\dfrac{(\text{line kv})^2}{(\text{connected transformer kva at source supplying fault})}$, using Fig. 23a.

2. $F \cdot X_A$ from $\dfrac{\text{connected transformer kva}}{\text{grounded transformer kva}}$, using Fig. 23b.

Note: Connected transformer kva and grounded transformer kva refer respectively to (1) total connected capacity, irrespective of transformer connections, and (2) transformer banks, which will pass zero sequence current.

3. X_L = apparent reactance (actually impedance) of equivalent power line mileage from source to fault. (Use actual mileage, if only one power circuit is involved), using Fig. 23c.
4. $X = F \cdot X_A + X_L$.
5. $Z = \sqrt{R^2 + X^2}$, where R = neutral and fault resistances, if any.
6. Short-circuit (fault) current I_f, using Fig. 23d.

Notes: If connected generator kva is more or less than the connected transformer kva by a factor of 2 or more, use connected generator kva supplying fault, for Figs. 23a and b.

If a reactance X_N is used in the grounding neutral, replace X with $X + X_N$.

Figures 23a, c, and d are based on the following assumptions:

1. A single 60-cycle generating station.
2. Radial transmission line.
3. Total connected transformer kva at station equals connected generator kva.
4. Line side neutrals of all transformer banks are grounded and will pass zero sequence current.
5. All transformer and generator capacities are so bussed as to be effective in supplying faulted line.
6. All transformers have 8 per cent reactance, based on kva rating of bank.
7. Representative average conductor spacings and sizes.
8. Earth resistivity of 100 meter-ohms (in the case of Fig. 23c).

The above procedure is not to be considered an exact method of short-circuit current calculation, since it usually provides only a rough determination, the method of symmetrical components being the proper one for more accurate results (see *Electrical Engineers Handbook*, Vol. 1, Electric Power, H. Pender and Wm. A. Del Mar, John Wiley & Sons).

Having determined the mutual impedance Z_M, and short-circuit current I_f, for any given situation and power line fault, the longitudinal induced voltage E_f, on the adjacent communication aerial open wires or cable, is given by the expression

$$E_f = I_f \cdot Z_M \cdot \eta$$

(Z_M from Fig. 22 or similar curves; η = shield factor.)

SHIELDING. Another important factor in determining the net coupling is the effect of grounded wires or other linear grounded metallic structures along the exposure. Voltages are induced in such grounded metallic structures in the same way as they are induced in telephone wires, and these voltages establish currents. The induced currents flowing in

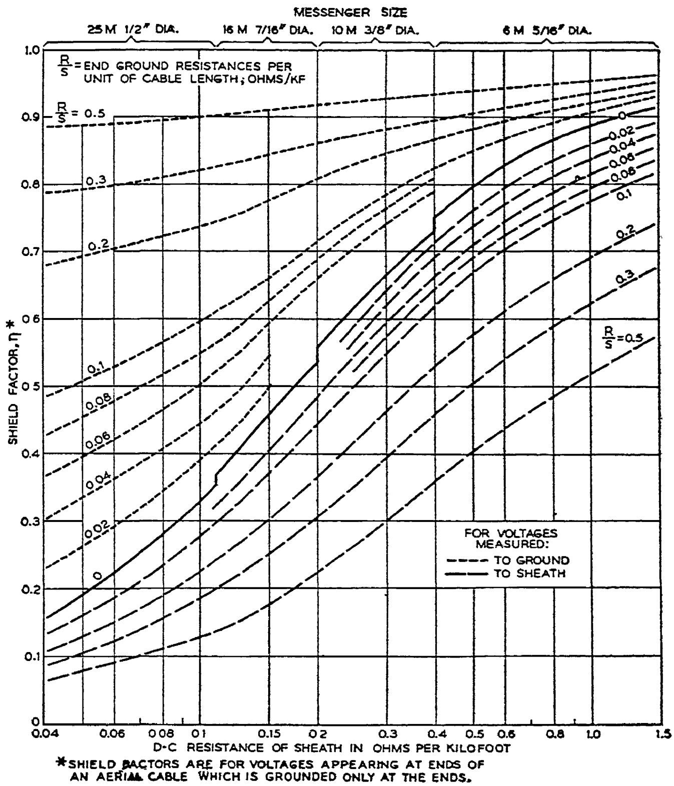


FIG. 24. Shield Factors for Aerial Lead Cable at 60 Cycles (Report 48, J.C.P.C.)

these structures set up magnetic fields, which generally oppose those from the power wires and reduce the induction in the telephone circuit. The effect of such currents in grounded structures is known as shielding.

The magnitude and phase relation of the current in a grounded conductor, and thus the shielding provided by it, depend on the impedance of the conductor with earth return. Hence, the shielding is increased, when the resistance of the conductor and its ground connections is reduced. Metallic cable sheaths comprise an important type of shielding conductor, as shown in Fig. 24 for aerial cables with plain lead sheath at 60 cycles; R is the sum of ground connection resistances at both ends of the cable sheath (in ohms) and S is the length of cable circuit between terminals or grounds (in kilofeet). With increasing end-ground resistance, the shield factors for voltages measured to ground increase, while the shield factors for the voltages to sheath (between sheath and conductors) decrease.

The earth resistivity was assumed to be 100 meter-ohms. Figure 25 shows shielding effects at 60 cycles from grounded telephone open wires.

In analyzing the distribution of induced voltage between a telephone circuit and ground, assume first that no protectors are operated. Under this condition, the voltages to ground on the telephone wires at various points are determined by the impedances between the wires and ground along the line and at central offices where equipment is connected to them. The voltage to ground at either end of the exposure is equal to the product of the longitudinal current and the impedance to ground seen looking away from the exposure at that end. In practice, the variety of impedance distributions encountered is almost infinite, and the corresponding voltage distributions vary over a wide range.

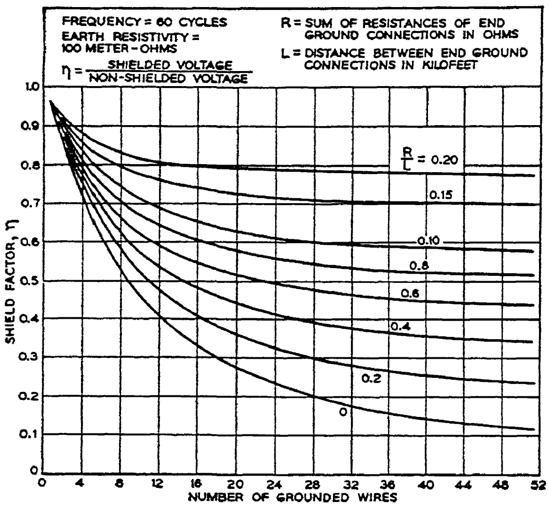


FIG. 25. Shielding Resulting from the Grounding of Conductors of an Open-wire Telephone Line (Courtesy Bell System)

If voltage to ground at any point where protectors are located exceeds the operating voltage of the protector, the protector operates and three things happen:

(*a*) The voltage to ground at the point of protector breakdown is reduced to a low value. This causes an increase in voltage across the protectors at the opposite end, and in most cases they also will operate.

(*b*) The operation of the protectors at the two ends completes a loop consisting of the telephone circuit and ground, so that the induced voltage will cause current to flow through both protectors.

(*c*) The voltages to ground on the circuits on which protectors have operated are changed and redistributed, and the voltages on the other telephone circuits are also changed and redistributed, owing to shielding.

All these effects take place within a very short time after the induced voltage is applied, so that, for all practical purposes, they can usually be considered instantaneous.

ACOUSTIC DISTURBANCE. Strictly the term "acoustic disturbance" should be used only with reference to the effect of an abnormally loud sound on a person subjected to it. The term has also come to designate a noise (usually transient) in a telephone receiver, the intensity of which is considerably higher than that of speech. It is produced by an excessive voltage across the terminals of the receiver.

Although induced voltages usually appear in equal magnitudes on the two sides of a circuit, the protector gaps on the two sides of the circuit discharge in an unsymmetrical manner, with the result that a voltage higher than normal appears across the circuit. When this occurs, a loud noise is produced in the receiver of a telephone set bridged across the circuit, causing acoustic disturbance. This may be produced by low-frequency induction, lightning, contacts between power and telephone circuits, and by other causes.

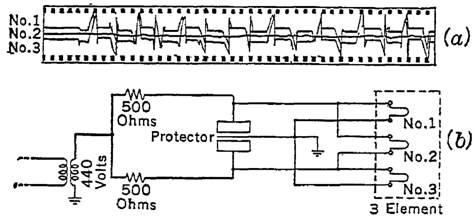


FIG. 26. Oscillograms Showing Protector Performance with Impressed Potentials across the Protector Blocks

Figure 26 shows oscillograph traces of voltages measured across operating protector blocks. Each outside trace shows the voltage across its related block to ground. It will be noted that the two traces are not identical. The middle trace shows the resulting voltage across the circuit. It is this voltage that may cause acoustic disturbance. The very jagged outline of these traces indicates that many frequencies are present.

PROBABILITY FACTORS. In the preceding discussion, a number of factors were mentioned which may vary between different occurrences in the same exposure. Among them are:

(*a*) The impedance in the faulted circuit. This varies with the location of faults, effective fault resistance, and other factors, with consequent effect on the magnitude of the fault current.

(b) The duration of the fault current. This varies with conditions that affect the speed of operation of the power-circuit interrupting devices.

(c) Longitudinal voltage, voltage to ground, and current through protectors. These may vary widely with small variations in the locations of faults when they occur within the exposure.

(d) Shielding due to the operation of protectors on telephone circuits. The number of protectors that operate may vary widely, depending on their characteristics and the magnitude of the induced voltage.

LOW-FREQUENCY INDUCTION CONTROL, though not usually required under normal operating conditions, must be planned for in advance where service may be affected by abnormal power-circuit conditions and where prompt action must be taken to restore normal operating conditions when abnormal conditions arise. Cooperative advance notices of construction and coordinative plans are essential in controlling the generally serious effects of low-frequency induction.

The low-frequency coordination of power and communication systems may be accomplished by (1) measures in the power system to limit the influence, (2) measures in the communication system to limit the susceptiveness, and (3) coordinated location of lines or other means to reduce coupling. In a specific situation, one measure may be sufficient or two or more measures may be required, depending on the conditions.

POWER SYSTEMS. Measures to reduce the inductive influence of power systems should be directed to limiting the magnitude of unbalanced currents and voltages, particularly under abnormal conditions, and to reducing the duration and frequency of occurrence of faults. Of such measures, some are concerned with questions of line and system design and must be incorporated in the construction plans and others may be added later, if found necessary as the result of operating experience.

Fault-resistive Design and Construction. Since lightning is a major cause of faults on power systems, the developments in methods of lightning protection have substantially aided the low-frequency coordination problem.

Fault-current Lmitation. Resistors or reactors in the neutrals of power systems provide a means of limiting the magnitude of the residual currents except when double faults occur.

Shielding. Ground wires on a power line, though they may increase the total residual current, provide shielding by reducing the strength of the external electric and magnetic fields. Under fault conditions, ground wires generally reduce the voltages induced in paralleling communication circuits. The effectiveness of such shielding depends on the impedance of the shielding conductor and its ground connections. Under favorable conditions, the induced voltage at 60 cycles may be reduced about 40 per cent by the use of one wire, and about 60 per cent by the use of two wires. However, such shield wires may increase the normal 60-cycle induced voltage, and this reaction may be important where ground-return telegraph circuits are involved.

High-speed Circuit Breakers and Relays. High-speed relay and switching systems have been developed that reduce the time duration of a power-line fault to the order of $^1/_5$ second or less under favorable conditions, thus tending to minimize the effects of induction. Because of the expense, high-speed switching can seldom be justified except on important high-voltage transmission systems.

Improvement in Balance. Low-frequency induction between power and communication lines is sometimes experienced under normal operating conditions. On grounded telegraph and signal lines, the trouble usually manifests itself by a chattering of telegraph instruments or by false signals. Sometimes power-line transpositions will aid in these situations.

COMMUNICATION SYSTEMS. In general, coordinative measures, applicable to the communication system, take the form of arrangements or devices for removing or counteracting the voltages, the voltage to ground, or the current through telephone protectors.

Relay Protectors. The short-circuiting relay (SCR) protector is designed for application to open-wire telephone lines which may be subjected to low-frequency induction of sufficient magnitude to warrant the costs of providing it. This device employs, for each open-wire circuit, a relay which short-circuits the usual protector blocks associated with the circuit and grounds the circuit. In one device of this type (Fig. 27), a master relay controls operation of all of the individual circuit relays, the master relay operating in about 0.01 second and the individual relays about 0.15 second later, after any one of the line protector blocks operates. The master relay operates when about 1 amp or more of 60-cycle current flows through the primary winding of the saturating transformer to ground. This produces a voltage across the secondary of the transformer, operating the master "J" type relay, causing the short-circuiting relays to operate and short-circuit each line and its associated protector blocks. The master relay can be adjusted to release on about 0.8 amp of line current. Thus, with operation of the pilot relay, all relays operate on about 0.5 amp direct current and ground all wires on the line at the locations where this equipment is installed. The equipment may be installed at any outside point on the open-wire line or in the central office, as desired. It is assembled in groups for application to 10 to 50 or more wires and, together with the dry-cell batteries for operating the circuit relays, is housed in standard cable terminal boxes.

Acoustic Disturbance Reducers. One of the most effective ways of reducing acoustic disturbance to operators is to shunt their receivers with a device that will have a high impedance at speech level and a low impedance at acoustic-disturbance level. One such device that has proved effective consists of oppositely poled copper oxide rectifier disks connected across the receiver. These have the property of greatly diminishing impedance with increasing voltage.

Shielding. Shielding on a telephone line may be effected by special grounded conductors, by working conductors, or by cable sheaths. Miscellaneous structures, such as pipe

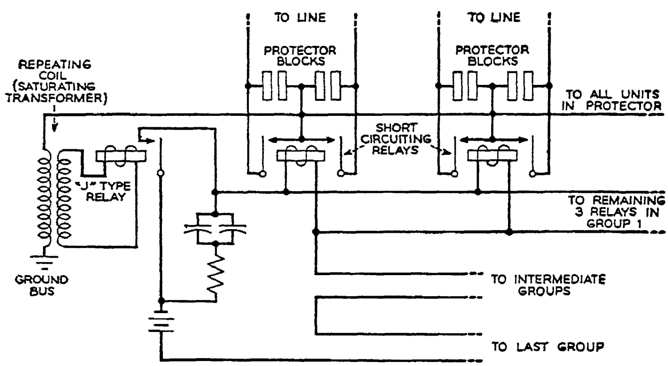


FIG. 27. A-c Type Short-circuiting Relay Protector Circuit—Multi-grounding (Report 41, J.C.P.C.)

lines or rails in the immediate vicinity of an exposure, also introduce shielding. A high-conductivity shield wire, well grounded at the ends of the exposure and at intermediate points, may reduce the induced voltage by as much as 40 per cent at 60 cycles.

The lead sheath of a 2 $^5/_8$-in.-diameter aerial telephone cable, if effectively grounded at the ends, reduces the voltage induced in the conductors within the cable by about 50 per cent at 60 cycles. The shielding secured from the sheath of an underground cable of this size is also about 50 per cent. The large number of conductors in a cable affords mutual shielding which varies from a negligible to a considerable amount (sometimes exceeding 95 per cent), depending upon many factors, important among which is the extent of the cable beyond the ends of the exposure and the grounding conditions on the circuits at their terminals. If two or more cables are close together through an exposure, each benefits by the shielding action of the others, so that the shielding increases with the number of cables.

If the lead sheath of the cable is surrounded by magnetic material, as by armoring or placing cable in iron pipe, the shielding may be largely increased. With a form of iron tape armored cable, shielding at 60 cycles may be 80 per cent or more, assuming effective grounding.

Other Measures. Sometimes drainage and neutralizing transformers may be of use. Drainage is achieved by grounding the midpoint of a coil connected between the two sides of the communication circuit, the coil being wound in such a way that it presents a low impedance to longitudinal currents and a high impedance to alternating currents in the metallic circuit. The voltage to ground on the communication circuit at the point where the drainage is connected is limited to the voltage drop over the impedance of the coil and ground connection.

The neutralizing transformer introduces into the exposed communication wires a voltage in opposition to the disturbing voltage, thereby partly neutralizing it, as described under "Noise Induction Mitigation" in article 32.

On account of the cost and the operating limitations that these measures impose on the use of carrier, d-c telegraph, and testing, they have only occasionally been employed in the commercial telephone plant. Neutralizing transformers have, however, been used on telegraph circuits and for special protection purposes on other types of communication and signal circuits.

BIBLIOGRAPHY

Principles and Practices

1. Reports of Joint General Committee of Edison Electric Institute and Bell Telephone System:
   (a) *Principles and Practices for the Inductive Coordination of Communication Systems*, Dec. 9, 1922.
   (b) *Principles and Practices for the Joint Use of Wood Poles by Supply and Communication Companies*, Feb. 15, 1926.
   (c) *Inductive Coordination—Allocation of Costs between Supply and Communication Companies*, Oct. 15, 1926.
2. Reports of Joint General Committee of American Railway Association (now Association of American Railroads) and Bell Telephone System:
   Inductive Coordination of Railway Electrical Supply Facilities and the Communication Facilities of the Bell System, Apr. 15, 1931, and Sept. 1, 1932.
3. Reports of Joint General Committee of the Edison Electric Institute and the Western Union Telegraph Company:
   Physical and Inductive Coordination of Electrical Supply and Communication Systems, July 22, 1935.
4. Reports of the Joint General Committee of the Edison Electric Institute and Bell Telephone System entitled *Physical Relations between Electrical Supply and Communication Systems* (E.E.I. Pub. M5).
5. *Year Book of the American Standards Association* (ASA).
6. Report of the Joint General Committee on Plant Coordination of the Edison Electric Institute and Bell Telephone System entitled *Joint Pole Practices for Supply and Communication Circuits* (E.E.I. Pub. M12), Oct. 29, 1945.
7. Safety Rules for the Installation and Maintenance of Electric Supply and Communication Lines (*National Bureau of Standards Handbook* H32, Sept. 23, 1941) (N.E.S. Code—5th Edition).
8. Discussion of the National Electrical Safety Code, Fifth Edition, Part 2, and Grounding Rules (*National Bureau of Standards Handbook* H39), July 15, 1944.
9. *Specification* 1-B-1 *for Communication Lines Crossing the Tracks of Railroads*, Association of American Railroads.

Technical Data

1. *Bibliography on Inductive Coordination by American Committee on Inductive Coordination* (1925). (List of articles and reports published before Jan. 1, 1925.)
2. Symposium on Coordination of Power and Telephone Plant, *A.I.E.E. Trans.*, Vol. 50, 437–478. Part I: Trends in Telephone and Power Practice as Affecting Coordination, by W. H. Harrison and A. E. Silver. Part II: Status of Joint Development and Research on Noise Frequency Induction, by H. L. Wills and O. B. Blackwell. Part III: Status of Joint Development and Research on Low-frequency Induction, by R. N. Conwell and H. S. Warren. Part IV: Status of Cooperative Work on Joint Use of Poles, by J. C. Martin and H. L. Huber. (These papers include references to the more important papers published before January 1931.)
3. Experimental Studies of Arcing Faults on a 75-kv Transmission Line, *A.I.E.E. Trans.*, Vol. 50, 1469.
4. Effects of Rectifiers on System Wave Shape, *A.I.E.E. Trans.*, Vol. 53, 54.
5. Petersen Coil Tests on 140-kv. System, *A.I.E.E.* Trans., Vol. 53, 63.
6. Iron Shielding for Telephone Cables, *A.I.E.E. Trans.*, Vol. 53, 274.
7. Overvoltages on Transmission Lines, *A.I.E.E. Trans.*, Vol. 53, 1301.
8. Joint Use of Poles with 6900-volt Lines, *A.I.E.E. Trans.*, Vol. 52, 890.
9. *Relay Handbook and Supplement*, N.E.L.A., 1931.
10. N.E.L.A. Publication 118, *Technical Theory of Inductive Coordination.*
11. N.E.L.A. Publication 144, *Some Features of Telephone and Telegraph Systems.*
12. N.E.L.A. Publication 147, *Cable Circuit Noise and Power Distribution Systems.*
13. N.E.L.A. Publication 153, *The Effects on Noise Frequency Induction of Grounded Neutral Generators.*
14. N.E.L.A. Publication 212, *Low-frequency Induction.*
15. N.E.L.A. Publication 239, *Generator Wave Shape.*
16. E.E.I. Publication A-13, *Power System Wave Shape.*
17. Low-frequency Power Induction and Its Effect on D-c Telegraph Operation, presented at annual meeting, Telegraph and Telephone Section, A.R.A., June 12, 1934.
18. Demonstration of Low-frequency Induction between Power and Telephone Circuits, also presented at above A.R.A. meeting.
19. Cooperation is Keynote in Rectifier Supply Coordination, by F. E. Sanford and V. G. Rettig, *Edison Elec. Inst. Bull.*, August 1935.
20. Earth Resistivity and Geological Structure, *A.I.E.E. Trans.*, Vol. 54, 1153.
21. Measurement of Telephone Noise and Power Wave Shape, *A.I.E.E. Trans.*, Vol. 54, 1307.
22. Engineering Reports of the Joint Subcommittees of the Edison Electric Institute and the Bell Telephone System:
    Vol. II, Report of Proj. Comm. No. 10, April 1932.
    Vol. II, Report 12, April 1932.
    Vol. II, Report 14, April 1932.
    Vol. IV, Report 34, January 1937.
    Vol. IV, Report 36, January 1937.
    Vol. IV, Report 38, January 1937.
    Vol. V, Report 41, January 1943.
    Vol. V, Report 44, January 1943.
    Vol. V, Report 45, January 1943.
    Vol. V, Report 46, January 1943.
    Vol. V, Report 48, January 1943.
23. A.I.E.E. Committee Report, *A.I.E.E. Trans.*, Vol. 65, 417.
24. Provisional Report 24 (Technical Report 2G-2), *A Study of Pole Strength in Jointly Used Poles*, of the Joint Subcommittee on Development and Research of the E.E.I. and B.T.S., Aug. 22, 1938.
25. A.I.E.E. Technical Report 47–82, *Joint Use of Pole Lines for Rural Power and Telephone Services*, by J. W. Campbell, L. W. Hill, L. M. Moore, and H. J. Scholz, December 1946.

SECTION 11

ELECTRICAL MEASUREMENTS

FREQUENCY MEASUREMENTS

By Warren A. Marrison

MEASUREMENT OF PRIMARY ELECTRICAL QUANTITIES

By J. G. Ferguson

WIRE LINE MEASUREMENT

By H. J. Fisher

ROUTINE MEASUREMENTS ON A-M AND F-M BROADCAST RECEIVERS

By W. O. Swinyard

WAVE ANALYSIS

By E. Peterson

MICROWAVE MEASUREMENTS

By E. W. Houghton

SIGNAL GENERATORS AND POWER MEASUREMENT

By F. J. Gaffney

ELECTRICAL MEASUREMENTS

FREQUENCY MEASUREMENTS

By Warren A. Marrison

1. DEFINITIONS

By *frequency* is meant the number of times any periodic phenomenon recurs in a standard unit of time. In general it is stated as the number of complete vibrations, or oscillations, or revolutions performed in 1 sec.

Occasionally other units of time are used, but usually they are specified so there is little or no ambiguity. Thus, we speak of revolutions per minute. Sometimes tuning forks are stamped DV or VS after the number designating the frequency. DV, which stands for "double vibrations," means frequency as defined above, that is, the number of complete periods per second. VS, which stands for "vibrations per second," designates twice the true frequency.

The *period* of a cyclic or periodic phenomenon is the time duration of one complete cycle. The rates of slow periodic phenomena are usually expressed in terms of the period. For example, the period of a "seconds" pendulum is 2 sec.

Frequency may also be designated as *angular velocity*, especially in expressions representing the rate as a trigonometric function of time. Thus in expressions of the type

$$I = I_0 \sin (\omega t + \alpha)$$

the constant, $\omega = 2\pi f$, may be considered as an angular velocity.

Frequency may be expressed in terms of *wavelength* when the velocity of wave propagation in a medium is assumed. This is most often used in discussions of electromagnetic radiation where the velocity is approximately that of light, $c = 2.998 \times 10^{10}$ cm per sec. The general relationship between frequency, velocity, and wavelength is

$$f \text{ (or } \nu) = \frac{\text{Velocity}}{\text{Wavelength}} = \frac{v}{\lambda}$$

A standard of frequency differs from most other working standards in that it is a *rate* and cannot be represented completely by a physical body that can be preserved. Although the frequency of a bar, or other simple shape, can be defined approximately for a given mode of vibration in terms of its dimensions, density, elasticity, and coupling to other bodies, the effects of these factors will not remain constant with as great accuracy as that to which the resultant frequency can be measured in terms of astronomical time, that is, in terms of the rate of the earth's rotation on its axis.

In the present system of measurements, the *primary standard* of frequency is the rate of the earth's rotation, since all measurements of frequency are referred directly or indirectly to the *second* of the cgs system which is by definition 1/86,400 of a mean solar day.

By the application of recently discovered atomic or molecular resonance phenomena to the measurement of frequency and time it may be found possible eventually to place all such measurements on a more nearly absolute basis, quite independent of the stability of masses of matter such as now comprise the control elements in reference standards of frequency and time, and independent of variations known to exist in astronomical time itself. Some of these atomic and molecular resonance phenomena are in the range for which continuous oscillations at ultra-high frequency may be produced by modern vacuum-tube means, indicating the possibility of making frequency comparisons of high precision throughout the entire range of continuous magnetic waves. The possibility of so using these resonance phenomena was discussed by Professor I. I. Rabi in a talk before the American Physical Society and the American Association of Physics Teachers in January, 1945.

The great importance of this impending development lies in the idea that the atomic and molecular resonance phenomena, being properties of independent elementary particles of matter, appear to be in certain cases completely independent of temperature, pressure, and other ambient conditions, which in various degrees affect the behavior of all existing practical standards of frequency and time.

Table 1. Some Useful Frequency Formulas

1. A-c generator with alternate N and S poles: $f = \dfrac{\text{No. of poles} \times \text{rps}}{2}$

   Inductor generator: $f = \text{No. of poles} \times \text{rps}$

2. Electrical resonant circuit with L and C: $f = \dfrac{1}{2\pi\sqrt{LC}}$

   Same with L, C, and R in series: $f = \dfrac{1}{2\pi\sqrt{LC - (R^2/4L^2)}}$

3. Frequency of electromagnetic radiation (λ = wavelength): $f = \dfrac{3.00 \times 10^{10}}{\lambda(\text{cm})}$

   Frequency of sound vibrations in any medium: $f = \dfrac{1}{\lambda}\sqrt{\dfrac{\text{Young's modulus}}{\text{Density}}}$

4. If a condenser is charged to voltage E and completely discharged f times a second into a current-measuring device: $f = \dfrac{\text{Mean current}}{C \times E}$

5. The period T of a simple pendulum with double amplitude $= 2\theta$: $T = 2\pi\sqrt{\dfrac{l}{g}}\left(1 + \dfrac{1}{4}\sin^2\dfrac{\theta}{2} \cdots\right)$

   The frequency for small amplitude: $f = \dfrac{1}{2\pi}\sqrt{\dfrac{g}{l}}$

   Conical pendulum (angle θ to vertical): $f = \dfrac{1}{2\pi}\sqrt{\dfrac{g}{l\cos\theta}}$

   Vertical pendulum (mass supported on spring): $f = \dfrac{1}{2\pi}\sqrt{\dfrac{\text{Stiffness}}{\text{Mass}}}$

   Torsion pendulum: $f = \dfrac{1}{2\pi}\sqrt{\dfrac{\pi\eta R^4}{2lI}}$
   where η, l, and R are the torsion modulus, length, and radius of the supporting member, and I is the moment of inertia of the mass.

6. Stretched string: $f = \dfrac{1}{2l}\sqrt{\dfrac{\text{Tension}}{\text{Mass per unit length}}}$
   Overtones are $2f$, $3f$, etc.

7. Uniform rod, free-free, longitudinal: $f = \dfrac{1}{2l}\sqrt{\dfrac{\text{Young's modulus}}{\text{Density}}}$
   Overtones are $2f$, $3f$, etc.

8. Uniform round rod, free-free, torsion: $f = \dfrac{1}{2l}\sqrt{\dfrac{\text{Torsion modulus}}{\text{Density}}}$
   Overtones are $2f$, $3f$, etc.

9. Long air column open at both ends: $f = \dfrac{1}{2l}\sqrt{\dfrac{\gamma\cdot\text{Pressure}}{\text{Density}}}$
   γ = ratio of the specific heats.
   Overtones are $2f$, $3f$, etc.

   Long air column open at one end: $f = \dfrac{1}{4l}\sqrt{\dfrac{\gamma\ \text{Pressure}}{\text{Density}}}$
   Overtones are $3f$, $5f$, etc.

10. Straight free-free bar in flexure: $f = \dfrac{9\pi k}{8l^2}\sqrt{\dfrac{\text{Young's modulus}}{\text{Density}}}$

    k is radius of gyration of section.
    Overtones are $(5/3)^{2f}$, $(7/3)^{2f}$, $(9/3)^{2f}$, etc.

    For flexural vibrations in general, tuning forks, reeds, etc., having uniform section: $f = K\dfrac{t}{l^2}\sqrt{\dfrac{\text{Young's modulus}}{\text{Density}}}$
    where K is a constant for a given shape and mode of vibration.

11. For any note in the equally tempered scale where n is the total number of semitones above (+) or below (−) middle C. (A = 440 cps.) $f_{\pm n} = 220 \times 2^{(3\pm n)/12}$

    Exact frequency ratios between successive notes in the major diatonic scale are: 9/8, 10/9, 16/15, 9/8, 10/9, 9/8, 16/15.

12. Where l is the distance between nodes on Lecher wires or in a coaxial conductor: $f \doteq \dfrac{3 \times 10^{10}}{2l}$

The most accurate reference standards of frequency in use at present are vacuum-tube oscillators whose frequencies are controlled by mechanical vibrators made of quartz crystal. A special advantage of the oscillator type of standard is that its output of constant-frequency current can be sent over suitable communication channels and used anywhere as a reference standard of frequency. Such standards are maintained continuously by the laboratories of the National Bureau of Standards, the Bell System, and many others in America and abroad.

Oscillators of somewhat less accuracy have been built employing tuning forks or bars of metal for the frequency-controlling element. Usually these are coupled to the vacuum-tube circuit through electromagnetic means, but many such oscillators have been built using magnetostriction and electrostatic attraction for the coupling means. Prior to the general use of quartz for precise control, such oscillators were used in most frequency standard installations. Now they are used chiefly where the advantage of direct low-frequency control outweighs the requirement for extreme accuracy.

Although, for the greatest accuracy, frequency is defined and measured in terms of astronomical time, preferably by means of a combined time and frequency reference standard, there are many cases in which a good estimate of its value can be made from a knowledge of means controlling it or responding to it. Table 1 contains a number of formulas that may be used for the approximate determination of frequency under a variety of conditions.

Frequency measurements find their chief applications in electrical communication where they are used in the study and adjustment of various oscillators and electrical networks such as filters and equalizers. This applies in varying importance from the lowest frequencies employed in d-c telegraph to the highest used in ultra-short-wave radio.

They are involved in the control of power frequencies and in certain time systems. They are used in measuring linear and angular velocity and acceleration, for the study of vibration in mechanical systems, and for the most precise determinations of electrical-circuit constants.

They are of ever-increasing importance in basic physical studies involving relations between astronomical time, the velocity of light, and resonance phenomena in atoms and molecules.

2. USING A TIME STANDARD

COUNTING. Since frequency is defined as the number of recurrences of a cyclic phenomenon per unit of time, the most direct method of its measurement, and at the same time the most precise, is to count the total number of cycles during a known time interval and to divide by the number of elapsed seconds. The only inaccuracies in this method are in determining the time interval and in estimating the number of cycles. If the time error can be assumed to be negligible, and if the frequency is constant, the accuracy may be increased to any extent by increasing the duration of the measurement.

Depending on the nature and the frequency of the phenomena to be measured, and the accuracy required, the instrumentation may vary over a very wide range. For low frequencies, and with relatively low accuracy, a stop watch may be used to count recurrences over intervals of a few seconds. For greater accuracy, and for frequencies too high for direct perception, some automatic registering or totalizing means is required. By means of a chronograph or oscillograph, a recording can be made of the recurrences during a specified time interval which may be analyzed at leisure. By various counting means a continuous total may be produced from which the frequency may be determined with great accuracy in any desired time interval.

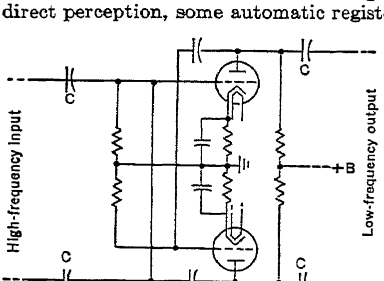


FIG. 1. Submultiple Generator of the Multivibrator Type. Capacitors C are for coupling only.

SYNCHRONOUS CLOCK. This in effect is the method used in calibrating the most accurate frequency standards of the quartz-controlled oscillator type. The actual method consists in operating a synchronous clock from the standard frequency source and in measuring the rate of the clock in terms of observatory time signals (astronomical time) by means of a chronograph. Since the accuracy of individual time signals may be somewhat better than 0.01 sec, the day-by-day accuracy of this method is of the order of 1 part in ten million. When the constancy of an oscillator justifies the use of a longer interval between checks, the accuracy of determination may be increased correspondingly. In this manner the rates of the best crystal oscillators may be determined in terms of astronomical time with an accuracy better than 1 part in a hundred million.

When accuracies of this order are involved, the variations in astronomical time itself should be taken into account, since changes in the rate of the earth in excess of 1 part in a hundred million have been observed from time to time. The largest such variation in recent years occurred in 1918 and amounted to about 1 part in thirty million.

Since the frequency of a crystal oscillator, best suited for use as a precise frequency standard, is too high to operate a synchronous motor directly, a circuit known as a submultiple generator is used to produce a frequency which is an exact submultiple of the original. Two or more stages may be required, depending on the ratio of the end frequencies. In one system that has operated continuously for over 10 years, the ratios 5 × 5 × 4 are used to operate 1000-cycle motors from a 100,000-cycle primary standard. Similar means may be used to obtain frequencies in any range convenient for distribution or for measurements which bear any exact rational relation to the primary control frequency. Two of the more important means for frequency subdivision are illustrated in Figs. 1 and 2. Although the former is most widely used, the latter has the advantage that no output is produced in the absence of an input.

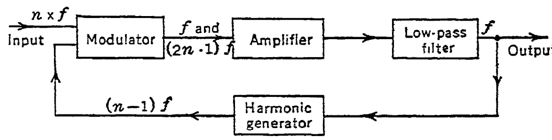


FIG. 2. Submultiple Generator of the Regenerative Modulator Type

3. USING A FREQUENCY STANDARD
(Frequency Comparison)

Methods for frequency comparison may be divided broadly into two main classes: those in which the actual number of cycles difference per unit of time may be determined between the unknown frequency and some simple exact fraction or multiple of the standard, and those in which some approximation is used that does not permit of actual counting. The first includes the various beat methods, the accuracy of which is limited only by the stability of the sources and the duration of observations. An accuracy of comparison of 1 part in 10^{10} may be obtained by some of these methods under good conditions. The latter includes (1) methods in which, for convenience and speed of operation, a calibrated interpolation oscillator is used to interpolate between standard frequency values, (2) methods in which circuit selectivity, or other non-synchronous means, are used to indicate frequency in some part of the system, and (3) methods in which the frequency of the source is in such a high range that the low-frequency beat contains irregularities and cannot be treated as a sine wave and used in any direct counting procedure. The accuracy of measurement in the last main class rarely exceeds 1 part in a million and may vary over a wide range, depending on the particular apparatus used and the skill of the observer.

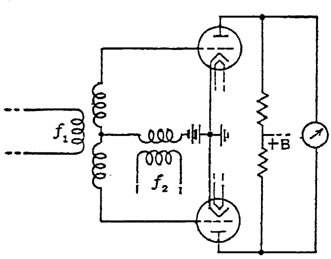


FIG. 3. Balanced Vacuum Tube Modulator for Observing Low-frequency Beats

ZERO BEAT. The simplest, most direct, and most accurate methods are those for comparing two frequencies which have nearly the same value. If the two sources are directed into a modulator so that the low-frequency second-order modulation produced goes through a d-c meter, the meter reading will vary periodically at a rate which is the difference between the two input frequencies. Thus the beats may be counted over a suitable interval and the number per second thus determined and added to or subtracted from the standard to obtain the unknown value. If the two frequencies are alike, there is no response, hence the term "zero beat." Figures 3 and 4 illustrate typical modulator circuits useful for this method.

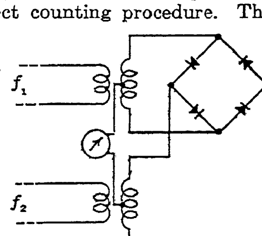


FIG. 4. Copper Oxide Modulator for Observing Low-frequency Beats

If a third-order modulator is used, such as silicon carbide, the greatest response is obtained with input frequencies near the ratio of 2 : 1. If higher-order modulation products are used, as by overloading vacuum tubes or other modulators, appreciable response can be obtained even when the input frequencies are related as $m : n$, provided that the product of the integers m and n is not too great. In this case beats are obtained between the nth harmonic of one input and the mth harmonic of the other. The numerical accuracy obtained is proportional to the ratio of the actual beat frequency observed to the frequency of the particular harmonic involved in the measurement. If the beat frequency obtained in this way is too high to observe directly, other means may be used to determine it.

The accuracy obtainable by this general method may be very high. For example, if the two frequencies, being in the ratio of nearly 1 : 1, are about 100,000 cycles, and if a beat of about 1 cycle in 10 seconds is obtained, an accuracy of only 10 per cent in the observed beat frequency corresponds to an accuracy of 1 part in ten million in the comparison of the two high frequencies. Highly stable sources may be compared readily by this means with an accuracy exceeding 1 part in 10^{10}.

OTHER BEAT METHODS. If the beat frequency obtained from a modulator is too high to observe directly, it may be measured by any other method suitable for the particular range encountered. For continuous comparisons the most precise method is to operate a synchronous clock from the beat frequency and to compute the frequency from its rate. A small percentage change in the rate of the undetermined high frequency will cause a much larger percentage change in the rate of the clock. For example, if the standard were 100,000 cycles and the undetermined frequency in the neighborhood of 100,100 cycles, the multiplying factor would be 1000.

If merely an indication of moderate accuracy is required, without integrating or recording, a commercial frequency meter, such as the vibrating-reed type, can be used giving a direct reading accurate to 1 cycle or a little better. The beat frequency must of course fall in the range of the particular instrument employed.

A very satisfactory direct-reading means for measuring the beat frequency in the range from one cycle to about 200 cycles per second is illustrated in Fig. 5. The modulator output operates a relay having a transfer contact combination. A condenser with capacitance C is charged to voltage E at the beat frequency f times per second. The meter then reads $f \times C \times E$, which is directly proportional to f. The factors C and E may be chosen in relation to the sensitivity of the meter so as to obtain a satisfactory scale reading. For very low frequencies, either a ballistic type of meter should be used, or a filter should be included in the meter circuit, in order to reduce meter fluctuation. Greater accuracy may be obtained, at the expense of the direct-reading feature, by balancing the voltage across a resistor, used instead of the meter, against a fraction of the voltage E, thus producing a null instrument. Both methods are suitable for continuous recording with commercial sensitive d-c recorders and are capable of high overall accuracy.

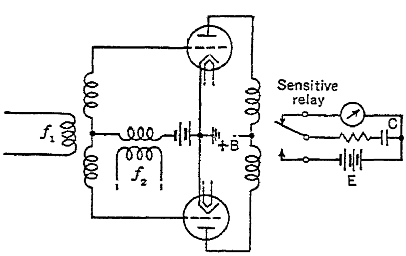


FIG. 5. Direct-reading Beat Frequency Indicator

Direct-reading instruments on the above principle have been built in which the relay is replaced by electron trigger tubes in order to extend the usable range of the beat frequency.

If the beat frequency is too high to measure accurately by any direct method, it can be determined in terms of a lower standard frequency by using a second modulating stage. This process can be carried out through a number of stages if desired so that, regardless of the value of the original undetermined frequency, the final beat is low enough to be measured accurately by some direct method. In one method that has been worked out in practice for measuring high frequencies, the frequency standard is made available in multiples of 1,000,000 cycles, 100,000 cycles, 10,000 cycles, etc., so that, by successive stages of modulation, a frequency determination is made in the range between 5000 and 30,000 kc in terms of a low-frequency beat and exact multiples of the several decade standards.

The frequency range over which this principle may be applied is limited only by the stability of the high-frequency oscillations. So long as the final beat frequency can be measured as a continuous sine wave, the limiting accuracy is about the same as for the zero-beat method. When, as is true at present of many ultra-high-frequency oscillators, the final beat signal contains such random variations that the cycles cannot be actually counted, some uncertainty enters even the best determinations of frequency. However, the accuracy may still be very high, because the beat frequency can be made such a very small part of the total range, thus reducing the necessity for great accuracy in its measurement.

STROBOSCOPIC METHODS. Stroboscopic methods are used for comparing the rate of one mechanical rotation or vibration with another or with the frequency of a fluctuating source of illumination. They are essentially low-frequency methods but have been used to measure speeds of rotation in excess of 10,000 rps. The accuracy of comparison is limited only by the constancy of the rates involved and the duration of a measurement, being in that respect like the beat methods just discussed.

In general a means is provided for permitting an observer to see an object periodically for very short intervals of time. If the object has a periodic structure like gear teeth or spokes, and if in the interval between glimpses it rotates a distance equal to one or any whole number of elements, it will appear to be stationary. If the glimpse frequency is not exactly equal to the apparent periodicity in the rotating structure, it will appear to move slowly, the amount and direction of motion corresponding to the relative rates. The time required for an apparent motion of the structure equal to one element space may be considered as the duration of one "beat" and treated as such in the comparison of rates. This method can be used in a large number of ways for measuring or comparing rotation speeds or for measuring frequency in terms of known rotation speeds, or vice versa.

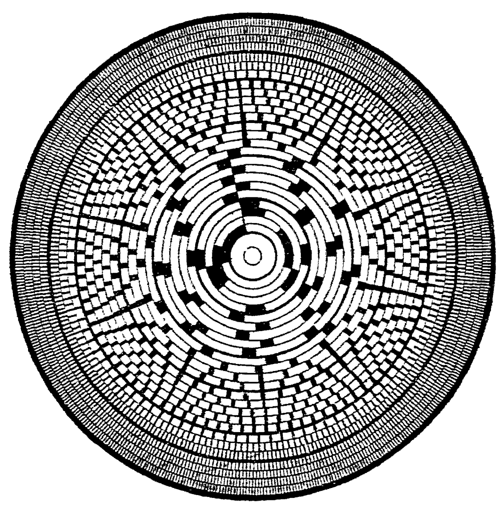


FIG. 6. Stroboscope Disk (Courtesy of General Radio Co.)

A simple apparatus for either case consists of a neon or other vapor lamp which can be flashed periodically by pulsating current, and a disk such as shown in Fig. 6 having concentric rows of black and white sectors of different periodicity, attached to a suitable rotating mechanism. If properly chosen, one row of sectors will appear stationary or nearly so in the intermittent illumination, and from the observation the ratio of the rotation speed to the flash frequency can be deduced readily and with great accuracy.

Similarly the frequency of any mechanical vibration of sufficient amplitude can be determined by observing the motion with a light flashed at a frequency that can be adjusted to the proper range. If the flash frequency is adjusted to the highest value that will give a single image of the vibrating part, that is the desired frequency, which can then be measured by electrical means or read from a calibrated dial.

CATHODE-RAY OSCILLOSCOPE. Probably the most useful of all laboratory apparatus for frequency comparison is the cathode-ray oscilloscope, now available in many compact and convenient forms. (See also Section 15, article 23.)

The oscilloscope, illustrated in Fig. 7, consists of a device for producing a stream of electrons which is directed toward a fluorescent screen within an evacuated tube, and of means for deflecting the electron beam in accordance with currents or voltages to be studied, thereby moving the luminous spot on the screen. At ordinary frequencies the motion of the spot is so rapid that its path is indicated by a continuous trace.

In the usual form of tube the deflections are obtained electrostatically by applying voltage to pairs of parallel electrodes between which the electron beam passes. When two voltage waves are to be compared they are connected to the two pairs of mutually perpendicular plates corresponding to the x and y axes in a cartesian coordinate system. The resulting motion of the spot on the screen is such that

$$x = ke_1 \sin (\omega_1 t + \varphi_1)$$
$$y = ke_2 \sin (\omega_2 t + \varphi_2)$$

where k is a constant for the tube and where e, ω, and φ are the voltage, angular velocity, and phase corresponding to the input waves.

The actual figure that is traced can be determined analytically in simple cases by eliminating t between the two equations. For example, if $e_1 = e_2$, $\omega_1 = \omega_2$, and $\varphi_2 - \varphi_1 = \pi/2$,

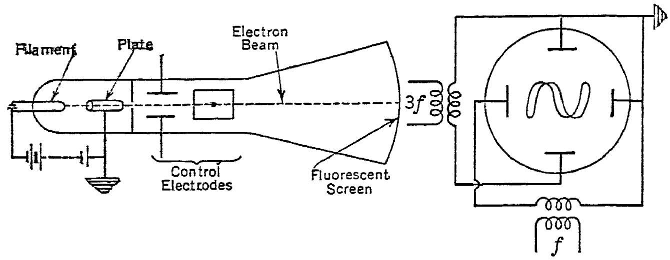


FIG. 7. Schematic of Cathode-ray Oscilloscope, Illustrating Electrostatic Method for Obtaining Lissajous Figures

we get $x^2 + y^2 = k^2e^2$, which is the equation of a circle. As the angle $(\varphi_2 - \varphi_1)$ changes we get various phases of an ellipse until, when it becomes 0 or π, it degenerates into a straight line inclined 45° to the axes. Thus when two waves are compared having equal amplitude and nearly equal frequencies, the pattern goes through a complete series of ellipses once for each cycle gained or lost by one frequency on the other. One such cycle

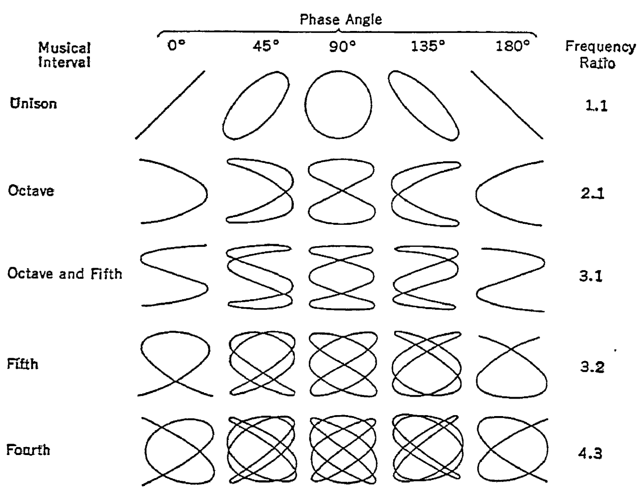


FIG. 8. Typical Lissajous Figures with Corresponding Frequency Ratios and Musical Intervals

of pattern changes corresponds to one "beat" between the input frequencies. Measuring the frequency of recurrence of such beats provides a convenient and extremely accurate comparison between the input frequencies.

Patterns formed in this way by frequencies that are equal or related as $m : n$, where m and n are integers, are known as Lissajous figures. Several such figures corresponding to simple frequency ratios and different phase angles are shown in Fig. 8.

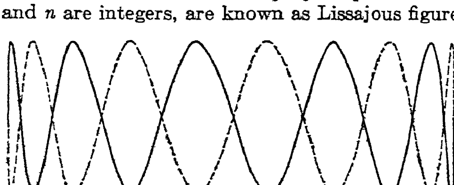


FIG. 9. Ten-to-one Lissajous Figure

When frequencies are to be compared which differ by a rather large ratio, say 10 : 1, a figure is obtained such as shown in Fig. 9. The figure looks like the projection of a sinusoidal trace of 10 cycles around a transparent cylinder. For illustration only, the drawing shows half of the figure solid, corresponding to the front side of the hypothetical cylinder, and the remaining half dotted. Actually, as viewed on the screen these two parts are indistinguishable in appearance. If the ratio departs very slightly from 10 : 1, the pattern moves as though the hypothetical cylinder rotated slowly at the rate of 1 complete revolution for every cycle gained or lost at the low frequency. Thus when the part shown solid in Fig. 9 moves to the right, the part shown dotted moves in the opposite direction.

In order to avoid the confusion of this double pattern, most cathode-ray oscilloscopes are now equipped with sawtooth wave sweep circuits, the frequency of which may be precisely controlled by an external standard. By this means the back pattern is eliminated and the whole visible pattern stands or moves as a unit.

Sometimes it is desirable to compare frequencies bearing a large ratio such as 100 : 1, or even 1000 : 1, in order to study small departures from such ratios, determined by other means, or to adjust the frequencies to those exact ratios. The patterns corresponding to such large ratios are too complex to be used in determining them directly, but by means of an auxiliary oscillator of intermediate frequency it is easy to obtain a nearly exact adjustment in two or three stages. For example, the 100 : 1 setting can be obtained by means of an auxiliary oscillator having 10 times the lower of the two frequencies concerned. As soon as the approximate adjustment is obtained between the end frequencies, the intermediate oscillator can be dispensed with and the final comparison carried out with the large ratio pattern. In such a procedure the horizontal gain of the oscilloscope may be made so large that only part of the complete figure appears on the screen. Because of the enlargement of the figure obtained in this way the accuracy of observation is materially improved. This is essentially a zero-beat method permitting extreme accuracy of comparison.

Frequencies over a very wide range may be compared by the cathode-ray-oscilloscope method, since many oscilloscope tubes will produce clear figures with inputs from the lowest frequencies up to many megacycles. The interpretation of multiline figures corresponding to fractional ratios and many special methods for using the oscilloscope are described in the references.

AUDIBLE METHODS. Audible methods are often useful, especially as a means for observing in one of the various beat methods. For example, if two frequencies which are nearly alike can be heard simultaneously, the beat frequency may be sensed as variations in loudness. The beats may be counted in order to determine the departure of one frequency from the other, or one of the frequencies can be adjusted to match the other by listening for zero beat. If one or both of the sound sources is rich in harmonics other ratios than 1 : 1 may be studied readily by this means.

When two audible frequencies are nearly alike and one source is movable in actual location, it is sometimes convenient to use the Doppler effect to determine which one is high. For example, if an observer holds a vibrating tuning fork in a stationary sound field of nearly the same frequency, slow beats will indicate the magnitude but not the sign of the frequency difference. However, if, while still vibrating, the fork is moved away from the observer and the beat frequency becomes lower (while moving), the fork frequency is higher than that of the sound field.

Often the musical pitch sense may be used to advantage in comparing the actual pitch of two tones one of which may be considered as standard. Since a musical semitone is approximately 6 per cent, it is evident that by this means alone the ratio of two tones in the musical range may be estimated to well within 5 per cent. As a direct measurement, this accuracy is insufficient for most purposes, but it is good enough to be very useful in making estimates of beat frequencies between two high-frequency sources. For example, a 5 per cent error in the 500-cycle beat between two frequencies, nominally 25 megacycles, corresponds to an error of only 1 part in a million in their comparison.

As an aid to this method it is convenient to keep in mind that the frequency ratios corresponding to the consecutive pairs of notes in the major diatonic scale are:

$$\frac{9}{8} \quad \frac{10}{9} \quad \frac{16}{15} \quad \frac{9}{8} \quad \frac{10}{9} \quad \frac{9}{8} \quad \frac{16}{15}$$

The product of all these together equals 2, that is, an *octave*. Also the product of the first four equals $^3/_2$, that is, the musical *fifth*, and similarly for the other recognized musical intervals.

Most keyboard musical instruments are tuned to the *equally tempered* scale in which the interval corresponding to all semitones is equal to $\sqrt[12]{2}$, and a whole tone equals two semitones. In Table 2 the actual frequencies are listed for a range including that of an 88 note piano with A above middle C equal to 440 vibrations per second. This is the most generally accepted standard of musical pitch. For the convenience of using round numbers, a pitch system is sometimes specified in which all the C's are powers of 2, middle C being 256 vibrations per second. The frequencies in this scale are about 2 per cent lower than in the accepted standard of musical pitch, corresponding to about a third of a semitone. The

Table 2. Equally Tempered Scale A = 440

	C_4–C_3	C_3–C_2	C_2–C_1	C_1–C	C–C′	C′–C^2	C^2–C^3	C^3–C^4
C	16.35	32.70	65.41	130.81	261.63	523.25	1046.5	2093.0
C♯	17.32	34.65	69.30	138.59	277.18	554.37	1108.7	2217.5
D	18.35	36.71	73.42	146.83	293.66	587.33	1174.7	2349.3
D♯	19.45	38.89	77.78	155.56	311.13	622.25	1244.5	2489.0
E	20.60	41.20	82.41	164.81	329.63	659.25	1318.5	2637.0
F	21.83	43.65	87.31	174.61	349.23	698.46	1396.9	2793.8
F♯	23.12	46.25	92.50	185.00	369.99	739.99	1480.0	2960.0
G	24.50	49.00	98.00	196.00	392.00	783.99	1568.0	3136.0
G♯	25.96	51.91	103.83	207.65	415.30	830.61	1661.2	3322.4
A	27.50	55.00	110.00	220.00	440.00	880.00	1760.0	3520.0
A♯	29.14	58.27	116.54	233.08	466.16	932.33	1864.7	3729.3
B	30.87	61.74	123.47	246.94	493.88	987.77	1975.6	3951.1
C	32.70	65.41	130.81	261.63	523.25	1046.50	2093.0	4186.0

latter pitch is used chiefly in physics and sometimes is known as physical pitch. International pitch is based on A = 435.

INTERPOLATION METHODS. Most measurements of frequency, apart from the intercomparison of standards and similar studies, can be made most expediently, and with sufficient accuracy, by the use of a calibrated interpolating oscillator in combination with some means such as just described for indicating exact frequency relationships. The principle is illustrated in Fig. 10, which may be modified or extended in numerous ways depending on the application. It is evident that other types of indicator than oscilloscopes could be used, and that by means of suitable switches, S_1 and S_2, only one indicator is needed for the simple example described in the following.

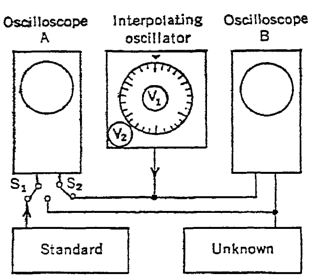


FIG. 10. Frequency Comparison with Interpolating Oscillator

The operation is best explained by an example. Let the standard be 100,000 cycles, and let us determine a frequency, say 3,564,000 cycles, by means of an interpolating oscillator stable in the range from 500,000 to 600,000 cycles and having a vernier dial, V_1, calibrated in this range. By a little experimenting one finds readily that an integral 6 : 1 pattern is observed on oscilloscope B when the interpolating oscillator is set a little below 600,000. That is the cue to calibrate the interpolating oscillator at 600,000 in terms of the standard by setting the dial to that exact value and adjusting vernier V_2 (which adjusts the frequency slightly, independently of V_1) until the ratio of the pattern on oscilloscope A is exactly 6 : 1. When the oscillator is again adjusted to give a 6 : 1 stationary pattern on oscilloscope B, the dial reading multiplied by 6 will be the frequency to be determined. In general it is best to so calibrate the interpolating oscillator that the two positions on the calibration dial are as near together as possible. When the oscilloscope is the indicator it is often desirable to use multiline figures in making one or both settings in order to accomplish this. For example, if the unknown frequency bore some simple ratio to 533,000 cycles, the three line pattern corresponding to the ratio $5\,{}^1/_3$ should be used in calibrating the interpolating oscillator. It is evident that the overall accuracy of the general method may be increased materially by this means especially if an auxiliary vernier dial is provided which is calibrated in a small percentage range and which can be used to interpolate between two close settings.

The method illustrated in Fig. 10 may be used in setting the "unknown" frequency at any value $P \times Q$ times the standard, where P and Q are integers or simple rational fractions. Using stationary figures on both oscilloscopes simultaneously, the accuracy of setting is very high. The "unknown" frequency can then be used as standard in a wide choice of frequencies to extend the range. Sometimes it is convenient to use more than one interpolating oscillator when it is necessary to cover a wide range of frequencies.

In particular this method can be used for extending the frequency of a standard upward for the purpose of measuring at ultra high frequencies in the decimeter and centimeter range. As yet, however, the oscilloscope cannot be used to obtain observable Lissajous figures at the highest frequencies because the stability of such oscillators is not yet good enough to produce stationary figures. This may be expected to come with time, however, and already cathode-ray tubes have been produced capable of resolving single traces at frequencies as high as 10,000 megacycles.

For ultra-high-frequency measurements in terms of a standard the usual method is to heterodyne a harmonic of a measurable source, such as the "unknown" of Fig. 10, with the high frequency by means of a crystal detector and either to estimate the frequency of the relatively low-frequency components obtained or to change the variable source until the detector output is as near to zero frequency as can be estimated. This does not have to be actually zero to be good; it should be remembered that, when measuring 10,000 megacycles, corresponding roughly to 3-cm waves, 10,000 cycles in the beat corresponds to only 1 part in a million in the overall measurement.

Standards of frequency of very great accuracy are made available by the National Bureau of Standards through continuous radio transmissions from station WWV. All the carrier frequencies, which are 2.5, 5, 10, 15, 20, 25, 30, and 35 megacycles, are regulated by the primary frequency standard of the bureau. Each carrier is modulated with seconds pulses and with the audio frequencies 440 and 4000 cycles, also of very high accuracy.

4. EMPLOYING CIRCUIT ELEMENT SELECTIVITY

When extreme accuracy is not a primary requirement, or when standard-frequency current is not available, it is often convenient to measure frequency in terms of the response of selective electrical networks or mechanical resonators.

METERS FOR POWER FREQUENCIES. Most of the commercial meters for indicating power frequency operate on one of three principles. Meters of the reed type employ a number of reeds tuned consecutively to slightly different values and loosely coupled to an electromagnet energized by the current to be measured. The reeds whose frequencies

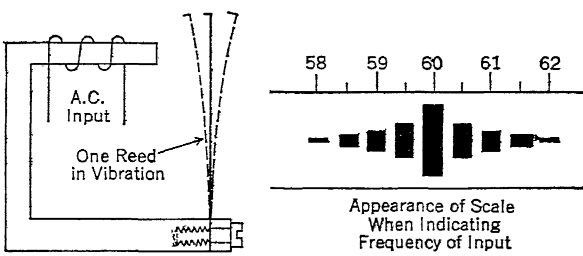


FIG. 11. Vibrating-reed Frequency Meter

correspond most nearly to the frequency of the input current vibrate at the largest amplitude. Usually the reeds are arranged in a row, as shown in Fig. 11, with the ends in line near a scale so that the frequency of the reed with greatest amplitude can be read off directly. Such meters are available in a considerable range of frequencies and are very useful for measuring low frequencies directly or for indicating beat frequencies in their range.

The Weston frequency meter, shown in Fig. 12, has a movable soft-iron armature free to move in the resultant field from two mutually perpendicular coils. When the frequency has some nominal value, the fields in the two coils are equal and the armature takes up a nominal position parallel to the resultant field. When the frequency changes, the ratio of currents in the coils changes, owing to the frequency selectivity of the input circuits, causing a shift in the resultant field and a corresponding movement of the armature.

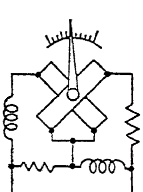


FIG. 12. Weston Frequency Meter

The induction frequency meter, shown in Fig. 13, consists essentially of two opposing induction voltmeter elements associated with one armature. The two motor elements are supplied through resistive and reactive circuits respectively so that the ratio of the effective currents varies with the applied frequency. In the type of meter shown, a circular disk is used and the indicating position is that for which a restoring spring just balances the resultant torque

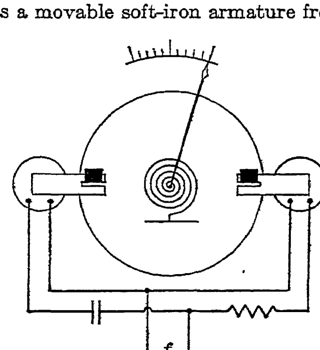


FIG. 13. Induction Frequency Meter

from the opposing drive elements. The deflection is therefore proportional to the frequency deviation from a nominal value.

In another type of this meter no restoring spring is used, but the armature is so shaped for one or both motor elements that the torque varies with angle of deflection as well as with input current. This permits an angular balance position to be obtained which does not depend upon the applied voltage or the reaction of a spring.

Some meters for power frequencies employ resonance to increase the sensitivity in a narrow frequency range. These and others are described in standard works on power meters.

BRIDGE METHODS. Various bridge methods may be used for measuring frequency. The one shown in Fig. 14 is typical. The bridge is balanced when

$$\left(R_2 + \frac{1}{j\omega C_2}\right)\left(\frac{1}{R_1} + j\omega C_1\right) = 1$$

With head phones, or other suitable null indicator, the frequency can be measured in terms of R_1, R_2, C_1, and C_2. As a frequency meter two of these elements can be made variable and calibrated to read input frequency at balance. For more detail about bridge measuring devices see below, article 12.

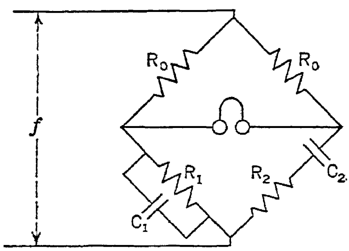


FIG. 14. Bridge for Measuring Frequency

THE MONOCHORD. This is a useful laboratory tool for the approximate determination of frequency in the lower and medium audible ranges. It consists of a steel wire under tension between movable bridges, the wire being coupled electromagnetically to the a-c input to be measured. The frequency of maximum response can be determined from the constants of the system, or a calibration can be made in terms of length or tension. Under some conditions an accuracy of 1 part in 1000 may be obtained with this means.

RADIO WAVEMETER. The tuned circuit wavemeter, Fig. 15, used extensively for radio-frequency measurements from 10 kc to 100,000 kc, consists primarily of a coil and condenser connected in a closed circuit and loosely coupled to a source to be measured. The circuit is tuned, usually by means of the condenser, until a resonance condition is indicated. From a scale on the condenser, previously calibrated by means of accurately known input frequencies, the frequency of any source in a limited range can be read off directly. In commercial wavemeters of this sort a set of coils is generally provided suitable for use in a number of adjacent and somewhat overlapping ranges.

The coupling from the source to be measured may be effected (1) through a low impedance, such as a low resistance, in series with the tuned circuit; (2) by loose magnetic coupling; or (3) by loose capacitance coupling as indicated in Fig. 15.

Resonance may be indicated (1) by a current-indicating instrument, such as a thermal galvanometer, in series with the tuned circuit; (2) by a voltage-indicating instrument, such as a crystal detector or diode in parallel with the reactive elements (if the source is modulated by audio frequency, head phones may provide the most convenient means for observing); (3) by the reaction on a power-indicating means associated with the source; (4) by a measure of power in a separate aperiodic circuit coupled to the tuned circuit but not directly to the source; and (5) by means of amplification followed by detection as in a simple radio receiver.

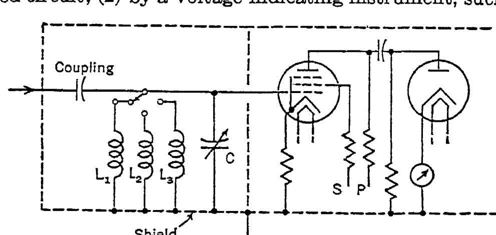


FIG. 15. Radio Wave Meter. Resonance type with high impedance resonance indicator.

Although all five methods are used, the fifth is preferable because of the vanishingly small effect of the indicating means on the Q of the tuned circuit and hence on the precision of observation. Used as indicated, an effective Q of 500 may be attained in a good part of the range, and with good equipment and careful procedure an accuracy of the order of 1 part in 10^4 may be achieved.

QUARTZ RESONATORS. Specific frequencies may be indicated with great accuracy by means of quartz resonators coupled loosely across the tuning elements of a resonant wavemeter. Owing to the relatively very much higher Q of the crystal, which is often in excess of 100,000, its response characteristic is confined to a correspondingly narrow and well-defined range of frequency. This response characteristic is superposed on the broader characteristic of the tuned circuit. Used in this way such resonators are valuable chiefly in fixing calibration points on the wavemeter. A variable input can be adjusted until the crystal response is indicated; then, with the same input frequency, the wavemeter condenser can be adjusted for maximum response to fix the calibration at the frequency of the crystal.

Quartz crystals may be used in a great variety of ways as single-frequency indicators. One interesting adaption, due to Giebe and Scheibe, consists of a series of resonators mounted in a gas at low pressure which may be coupled electrostatically to a circuit to be tested. If a frequency is applied which corresponds to that of one of the resonators, the large potential gradients in the neighborhood of that resonator due to its resonance will cause a luminous discharge which serves as indicator. This method is somewhat analogous to the reed indicator but is applicable to frequencies up to a million cycles per second or more and of course is inherently much more accurate.

CAVITY RESONATORS. The most convenient and generally satisfactory means for measuring frequencies in the region from 200 to 30,000 megacycles has been cavity resonators, which have been developed in a variety of forms for different frequency ranges and for different methods of use. A cavity is primarily an almost completely enclosed space in a rigid piece of metal, with openings only for coupling electromagnetic energy from wave guides or coaxial conductors, and usually also for the operation of a plunger for frequency adjustment.

The cavity itself in its lowest-frequency modes may be considered a familiar tuned circuit in which the *capacitance* is formed by opposite sides, sometimes the two flat ends of a cylinder, and connected by a continuous array of single-turn coils forming the cylinder wall. It is evident that for a moderate-sized cavity the effective capacitance and inductance are both very small and hence the frequency is very high. Since there are no radiation losses, and since the interior losses may be kept small, the Q factor may be very high. For silver-plated fixed cavities the Q may exceed 20,000; for adjustable cavities it is somewhat less.

Cavities are resonant elements and as such may be used as selective transmission devices indicating a *maximum* of transmitted energy into a detector, or as selective absorption devices indicating a *minimum* in a high-impedance source. In either case the actual detection is most easily accomplished by a crystal detector which may actuate a d-c meter, or, if the source is modulated, a set of head phones may be used, with amplification if necessary. The setting accuracy may be as high as 1 part in 10^5.

The design, construction, and use of cavity resonators are discussed at length in Section 7 and in references.

TRANSMISSION LINES. Very high frequencies can be measured with fair accuracy by a study of standing waves in transmission lines, the three usual types being Lecher wires, coaxial conductors, and wave guides. The simplest method involves a movable short-circuiting conductor which can be moved along the line by external control, and the observation of successive minima due to reaction on the source of high-frequency energy fed into the line. The actual distance between successive positions of the short-circuiting conductor which causes such minima determines a half wavelength for the particular line used. In the case of Lecher wires (parallel wires a few centimeters apart) and coaxial lines, the frequency is given approximately by

$$f = \frac{3 \times 10^{10}}{2l}$$

where l equals the distance between resonance positions. A calibrated instrument based on this method may be accurate to about 0.1 per cent.

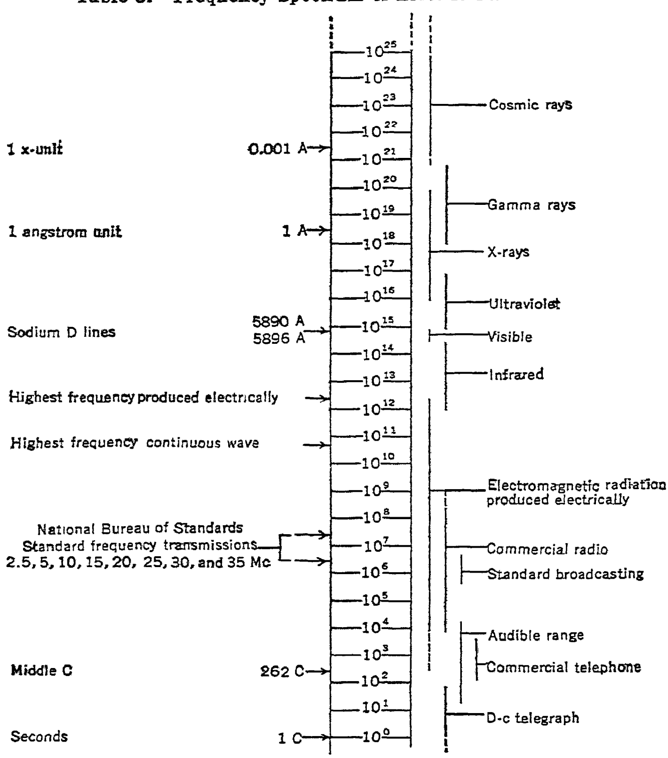
Although Lecher wires are convenient for purposes of demonstration, a low-resistance lamp on the short-circuiting rider being a suitable indicator of resonance, the attainable accuracy is limited by large radiation losses, especially at high frequencies. Wave guides are inconvenient for this purpose on account of the wide variety of modes that may give false indications unless special precautions are taken to suppress them. High-frequency transmission lines are discussed in Section 10, and some of the accompanying references deal with their use as frequency-measuring devices.

THE ECHELETTE GRATING. A method, closely analogous to the familiar optical grating, has been developed for establishing the frequency of the shortest electromagnetic radiations producable experimentally. The method consists in directing the radiation in question toward an array of stepped metal reflectors, like the slats in a venetian blind, and in observing the angle of reflection for which the greatest intensity of reflected energy is obtained. From the spacing of the reflectors, and the measured angles for maximum energy, the wavelength may be computed with considerable accuracy as with the optical grating. This method of measurement, in fact, was an important step in establishing the continuity of the electromagnetic spectrum between light produced by thermal excitation and electric waves produced by purely electrical means.

5. ELECTROMAGNETIC PHENOMENA

Table 3 shows in a graphical manner the relation between the frequencies of various observed electromagnetic phenomena. Where any considerable uncertainty exists as to the extent of a range due to ambiguity of definition or due to disagreement among sources, the uncertainty is indicated by dotted extensions.

Table 3. Frequency Spectrum of Electrical Phenomena



The lower end of the chart, and extending a little beyond 3×10^{10}, includes the entire range over which continuous alternating current can be produced at the present time. This is also the entire range over which frequencies can be measured or compared directly. Damped waves having frequencies up to 3×10^{12} have been produced by purely electrical methods. In this region, which overlaps the infrared spectrum, the frequencies can be determined by either electrical resonance methods or by optical interference methods.

From 10^{12} upwards the terms on the right refer to electromagnetic waves originating in hot or otherwise luminous bodies, in excited atoms, or in atom nuclei. Between 10^{12} and 10^{20} the frequencies are calculated from measured wavelengths; from 10^{20} upward they are calculated from measured photon energies. In certain ranges either of two names is applicable, as indicated by the overlapping grouping.

It is interesting to note that, by extending the frequency spectrum downward, frequencies of recurring astronomic phenomena are encountered that are about as far removed from one cycle as are the highest frequencies noted. Thus the frequency of rotation of the earth on its axis is 1.16×10^{-5}; the frequency of revolution of the earth around the sun is 3.17×10^{-8}. The frequency of rotation of the equinoxes around the ecliptic is

1.2×10^{-12}. The frequency of rotation of Andromeda Nebula is estimated (Jeans) to be 1.7×10^{-15}, and it is perhaps reasonable to suppose that frequencies of periodic phenomena involving interactions between the nebulae may be many orders smaller. Thus the frequencies to which the senses respond most readily are, on a logarithmic scale, about midway between the extremes of which we are aware.

BIBLIOGRAPHY

Books

Jansky, Cyril M., *Electrical Meters*, McGraw-Hill, 1917.
Pierce, George W., *Electric Oscillations and Electric Waves*, McGraw-Hill, 1920.
Brown, Hugh A., *Radio Frequency Electrical Measurements*, McGraw-Hill, 1931.
Hund, August, *High Frequency Measurements*, McGraw-Hill, 1933.
Brainerd, J. G., Kochler, Glen, Reich, Herbert J., and Woodruff, L. F., *Ultra High Frequency Techniques*, D. Van Nostrand, 1942.
Montgomery, C. G., *Technique of Microwave Measurements* (Rad. Lab. Series 11), McGraw-Hill, 1947.

Articles

Cady, W. G., The Piezoelectric Resonator, *Proc. I.R.E.*, Vol. 10, No. 2, 83–114 (April, 1922).
Nicols, E. F., and Tear, J. D., Short Electric Waves, *Phys. Rev.*, Vol. 21, Series II, 587–610 (June 1923).
Pierce, George W., Piezoelectric Crystal Resonators and Crystal Oscillators Applied to the Precision Calibration of Wave Meters, *Proc. Am. Acad. Arts and Sci.*, Vol. 59, No. 4, 81–106 (October 1923).
Hund, August, Theory of Determination of Ultra-radio Frequencies by Standing Waves on Wires, *Scientific Papers of Bureau of Standards* No. 491 (June 23, 1924).
Giebe, E., and Scheibe, A., Leuchtende piezoelektrische Resonatoren als Hochfrequenznormale, *E.T.Z.*, Vol. 47, 380–385 (April 1926).
Clapp, J. K., Universal Frequency Standardization from a Single Frequency Standard, *J.O.S.A.* and *R.S.I.*, Vol. 15, No. 1, 25–47 (July 1927).
Rasmussen, Frederick J., Frequency Measurements with the Cathode-ray Oscillograph, *J. A.I.E.E.*, Vol. 46, 3–12 (July 1927).
Allen, George E., The Accuracy of the Monochord as a Measurer of Frequency, *Phil. Mag.*, Vol. 4, Series 7, 1324–1337 (December 1927).
Ferguson, J. G., and Bartlett, B. W., The Measurement of Capacitance in Terms of Resistance and Frequency, *Bell Sys. Tech. J.*, Vol. 7, 420–437 (July 1928).
Pierce, George W., Magnetostriction Oscillators, *Proc. I.R.E.*, Vol. 17, 42–88 (January 1929).
Marrison, W. A., High Precision Standard of Frequency, *Bell Sys. Tech. J.*, July 1929, pp. 493–514, *Proc. I.R.E.*, July 1929, pp. 1103–1122.
Marrison, W. A., A Method for Estimating Audible Frequencies, *Bell. Lab. Rec.*, Vol. 8, No. 4, 178–182 (December 1929).
Polkinghorn, F. A., and Roetken, A. A., A Device for the Precise Measurement of High Frequencies, *Proc. I.R.E.*, Vol. 19, 937–948 (June 1931).
Johnson, J. B., The Cathode-ray Oscillograph, *J. Franklin Inst.*, Vol. 212, No. 6, 687–717 (December 1931).
Loomis, Alfred L., and Marrison, W. A., Modern Developments in Precision Clocks, *Trans. A.I.E.E.*, June 1932, pp. 527–537.
Peterson, H. O., and Braaten, A. M., The Precision Frequency Measuring System of R.C.A. Communications, Inc., *Proc. I.R.E.*, Vol. 20, No. 6, 941–956 (June 1932).
Mickey, L., and Martin, A. D., Development of Standard Frequency Transmitting Sets, *Bur. Standards J. Research*, Research Paper 630 (January 1934).
Dye, D. W., and Essen, L., Valve Maintained Tuning Fork as Primary Standard of Frequency, *Proc. Royal Soc.*, January 1934, pp. 285–306.
Forbes, H. C., and Zaugbaum, F., Frequency, Time Control with Telephone Aid, *Elec. World*, Jan. 20, 1934, p. 117.
Scheibe, A., und Adelsberger, U., Die technischen Einrichtungen der Quarzuhren der Physikalisch-Technischen Reichsanstalt, *Hoch frequenztechnik und Elektroakustik*, February 1934, Band 43, Heft 2, pp. 37–47.
Hazen, Grace, and Kenyon, Frieda, Primary Radio-frequency Standardization by the Use of the Cathode-ray Oscillograph, *Scientific Papers of Bureau of Standards*, No. 489 (May 1934).
Williams, Emrys, Audio-frequency Measurement by the Electrically Excited Monochord, *Proc. I.R.E.*, Vol. 22, No. 6, 794–804 (June 1934).
Cleeton, Claud E., Grating Theory and Study of the Magnetostatic Oscillator Frequency, *Physics*, June 1935, pp. 207–209.
Cleeton, C. E., and Williams, N. H., The Shortest Continuous Radio Waves, *Phys. Rev.*, Vol. 50, 1091 (December 1936).
Essen, L., and Gordon-Smith, A. C., The Measurement of Frequencies in the Range 100 mc/s to 10,000 mc/s, *J. I.E.E.*, Part III, Vol. 92, No. 20, 291–295 (December 1945).
Jones, William J., Types and Applications of Microwave Frequency Meters, *Radio*, January 1946, pp. 29–34.
Clayton, R. J., Houldin, J. E., Lamont, H. R. L., and Wilkshaw, W. E., Radio Measurements in the Decimeter and Centimeter Wave Bands, *J. I.E.E.*, Part III, Vol. 93, No. 22, 97–117 (March 1946).
Lee, Gordon M., A Three-beam Oscillograph for Recording at Frequencies up to 10,000 Megacycles, *Proc. I.R.E.*, and *Waves and Electrons*, Vol. 34, No. 3, 121W–129W (March 1946).
Essen, L., Cavity Resonator Wave Meters, *Wireless Engineer*, Vol. 23, No. 272, 126–132 (May 1946).
Booth, C. F., and Laver, F. J. M., A Standard of Frequency and Its Applications, *J. I.E.E.*, Part III, Vol. 93, No. 24, 223–236 (July 1946).
Gaffney, F. J., Microwave Measurements and Test Equipment, *Proc. I.R.E.*, and *Waves and Electrons*, Vol. 34, No. 10, 775–793 (October 1946).
Lafferty, J. M., A Millimeter-wave Reflex Oscillator, *J. Applied Phys.*, Vol. 17, No. 12, 1061–1066 (December 1946).
Marrison, W. A., The Evolution of the Quartz Crystal Clock, *B.S.T.J.*, Vol. 27, 510–588 (July 1948).

MEASUREMENT OF PRIMARY ELECTRICAL QUANTITIES (CURRENT, VOLTAGE, RESISTANCE, CAPACITANCE, AND INDUCTANCE)

By J. G. Ferguson

The primary quantities of interest in the communications field may be divided into two classes: first, current and voltage; and second, the so-called circuit constants, resistance, capacitance, and inductance. Power is usually obtained from the measurement of voltage or current, and resistance. Frequency is covered under Frequency Measurements, pp. 11-1 to 11-15.

Values of current and voltage generally do not require very accurate measurement. On the other hand, very severe requirements are justified economically in the design and measurement of circuits and their component parts if they result in an increase in the number of channels which can be made available from a given physical circuit. These requirements are specified for the most part in terms of the three circuit constants, resistance, capacitance, and inductance.

The frequencies of interest in communication circuits range from a few cycles per second to the super high frequencies used in radio transmission, whereas the frequency range of major interest from the standpoint of the measurement of primary circuit constants is from about 30 cycles to about 100 megacycles.

6. MEASUREMENT OF CURRENT

The values of power of interest have a lower limit of about 10^{-16} watt determined by the necessity of keeping a level above resistance noise, and an upper limit of about 1 watt in wire transmission, determined by the necessity of avoiding excessive cross-talk to other circuits. The upper limit, of course, is considerably higher in certain applications such as radio transmitters. The currents corresponding to these power limits for the impedance ranges encountered range from about 0.1 amp down to fractions of a microampere. The three features most desired in an instrument to measure such currents are: an impedance practically non-reactive and independent of frequency and current level, a method of operation which furnishes effective values, and high ratio of response to input power.

TYPES OF INSTRUMENT. The dynamometer type and the magnetic-vane type measure effective values of current but require high input power. They are seldom used except at power frequencies. Practically all a-c measurements of current use as an indicator a d-c instrument of the moving coil-permanent magnet type. The problem then reduces to a means of transforming the alternating current to direct current.

THE THERMOCOUPLE TYPE. This has an impedance which is practically a pure resistance at all frequencies, and it measures effective values of current. It is the most accurate method for the measurement of small currents because it can be calibrated with reversed direct current. Its principal disadvantages are that, irrespective of the current measured, the output power is low, about 1 microwatt, and it will not stand heavy overloads. Enclosing the couple in a vacuum improves its speed and sensitivity and reduces temperature errors. The usual type has a heater in direct contact with the junction. This gives a maximum sensitivity and speed and is satisfactory for all but the highest frequencies. At very high frequencies the direct contact between the a-c and d-c circuits introduces objectionable couplings between the instrument and other parts of the circuit. A type having the couple insulated from the heater is to be preferred above about a million cycles in spite of a somewhat slower speed and slightly lower efficiency.

Typical vacuum couples of either the insulated or contact type have a couple resistance of about 10 ohms and are designed to work into a microammeter having a resistance of about the same value and a full-scale deflection of 200 to 300 μa. Table 1 gives essential information for instruments of this type having various ranges. Their accuracy is usually

Table 1. Typical Ranges and Resistances of Thermocouple Instruments

Range, milli-amperes	Input Resistance, ohms	Range, milli-amperes	Input Resistance, ohms	Range, milli-amperes	Input Resistance, ohms
250	0.5	25	10	5.0	100
100	2.0	10	35	2.0	600
50	4.0	7	50	1.5	1000

about 1 per cent of full-scale reading when used with the couple with which they have been calibrated.

Multirange milliammeters are made with self-contained shunts. The accuracy of such instruments at high frequencies is usually limited by the shunts.

Contact Rectifier Type. Contact rectifiers used in conjunction with d-c meters are the most rugged and efficient of the various low current instruments. The most common type uses the copper vs. copper-oxide disk. These instruments have the disadvantage, common to all rectifiers, that they introduce modulation due to instantaneous variation of the input resistance with current. This may be reduced considerably by the common arrangement of four units in bridge form, or an equivalent arrangement with two units and a center tapped transformer. They usually measure more nearly average than effective values, thus giving a wave-shape error. Their principal disadvantages lie in the variation of the rectification properties with time and temperature and the fact that the input impedance varies with current and has a large capacitive component which acts as a shunt across the rectifier and renders their use above about 10,000 cycles of little value, unless some compensation is provided. With proper compensation their frequency range may be extended considerably. For higher frequencies silicon or germanium crystal rectifiers may be used.

Since different ranges can be obtained by changing either the size of the disks or the sensitivity of the d-c meter, a very wide current range is available. It is possible to obtain meters with a full scale as low as 100 μa, but their resistance is high, about 3000 ohms. This makes them suitable for voltmeters, which can be obtained with ranges as low as 0.5 volt with a resistance of 3000 ohms per volt.

The accuracy of this type of instrument is usually limited to about 5 per cent of full scale, and errors in excess of this may occur with extremes of temperature and wave shape.

Two-element vacuum-tube rectifiers have characteristics somewhat similar to contact rectifiers. They are satisfactory up to much higher frequencies but have the disadvantage of requiring auxiliary power. They are discussed further under voltmeters.

MEASUREMENT OF HIGH CURRENT VALUES. These measurements may be made by means of low-range meters in conjunction with current transformers or shunts. Shunts are preferred for high frequencies. They should be designed to have impedance characteristics similar to those of the meter, particularly when used at high frequencies, so that the current will divide in the same proportion at all frequencies.

MEASUREMENT OF LOW CURRENT VALUES. Measurement of current values below the range of even the most sensitive instruments can be made after amplification by comparison with a known current obtained by attenuating a measured current a definite amount.

7. MEASUREMENT OF VOLTAGE

With the high impedances almost inevitable in an instrument for measuring low current values, there is no marked distinction between the measurement of current and voltage; thus all the methods discussed for current measurement will also measure voltage. However, care must be taken to avoid errors due to impedance change caused by series reactance in the leads at high frequencies. Static voltmeters are also common. They measure effective values, but their low sensitivity and high input capacitance limit their use to comparatively high voltages and low frequencies. In addition the following are available.

VACUUM-TUBE RECTIFIERS. This type provides the most satisfactory means of voltage measurement. Most vacuum-tube voltmeters employ diode rectifiers with high-resistance loads, rather than triodes, since the indication is less dependent on the tube characteristics, and because plate-voltage variations, which are a source of error in the triode, are eliminated.

To measure low voltages the rectifier is either preceded by an a-c amplifier or followed by a d-c amplifier. The frequency range and sensitivity of the former are limited by the problem of obtaining broad-band stabilized amplification. Instruments are available in multiple-range models with sensitivities in the order of 1 millivolt, covering the frequency range of 20 cycles to 5 megacycles. The input impedance is in the order of 0.5 megohm shunted by a few micro-microfarads. Indicated voltages are usually average values.

The sensitivity of diode voltmeters is limited by the diode characteristic, and the frequency is limited principally by lead inductance. To reduce this error, the diode is usually encased in a small probe connected to the set by a cable. Instruments are available in multiple-scale models with sensitivities of 0.1 volt covering the frequency range of 20 cycles to 100 megacycles. Input impedance is usually 1 megohm or more at low frequency but is limited at high frequency by the capacitance and conductance of the diode, which may be as high as 7 $\mu\mu$f and 10 micromhos at 10 megacycles. Indicated voltages are usually half wave peak values.

The accuracy of vacuum-tube voltmeters is 2 to 5 per cent, depending on wave shape and frequency.

For frequencies above the range of the diode, materials of high temperature coefficient of resistance are used in which the voltage is determined by the resistance change due to heating. Owing to their small size they may be made independent of frequency up to several thousand megacycles.

8. RESISTANCE STANDARDS

For the accurate measurement of circuit constants, precision standards are required. They form the most important part of the measuring circuits and therefore require full consideration. They consist of resistance, capacitance, and inductance. The following discussion is limited to their use as standards. Further general information can be found in Section 3.

Resistance standards are usually wire wound, but high resistances are also made by coating a thin film, usually of carbon, on an insulating form. These films have a high temperature coefficient of resistance, about -0.03 per cent per degree centigrade for carbon, and cannot be adjusted to very close limits. They are fairly stable with time if hermetically sealed, and the change with frequency is small for values below 10,000 ohms. They are cheaper than wire wound, especially for high resistance values.

The types of wire used most generally for resistance standards are the copper-nickel alloy Constantan and the copper-nickel-manganese alloy Manganin. Both materials have very low temperature coefficients, the variation in resistance over the temperature range of ordinary use being well below 0.005 per cent. Constantan is more easily soldered, and its resistance is more stable with time. Manganin is preferred for d-c standards on account of its lower thermoelectric power to copper, but this is a minor consideration for a-c use. Both alloys have a specific resistance of about 30 times that of copper. Other alloys containing chromium or iron have higher specific resistances but are not generally used on account of their higher permeabilities and larger temperature coefficients.

EFFECT OF HUMIDITY. Humidity affects the distributed capacitance of resistance windings unless they are sealed or impregnated. The impregnating material is usually a thin solution of shellac or a wax, such as paraffin. Impregnation also serves to prevent change in shape of the support or spool, but it is desirable to use as a support a material impervious to moisture, such as a suitable plastic, ceramic, or glass. The woven type of wire standard is the most independent of changes in the form.

PHASE ANGLE. The phase angle should be small. It is due to inductance, distributed capacitance, and capacitance between the terminals which is not due to the presence of the winding. Inductance can be reduced by reducing the size of wire and by winding the coil so that adjacent turns have opposite directions of winding and are spaced as close together as possible. Minimum spacing is determined by the thickness of insulation. Therefore, for a given type of wire and insulation there is a definite minimum obtainable ratio of inductance to resistance for any given wire size, this value being reduced as the wire size is reduced. Distributed capacitance can be reduced by winding so that turns which are adjacent physically are consecutive electrically. By this means the distributed capacitance can be reduced below the remaining capacitance between the terminals. Both types of capacitance can be decreased by reducing the physical size, that is, by reducing the wire size. Capacitance and inductance, when both are present in a resistance, have a compensating effect. For the small reactances present in resistance standards, the compensation will be almost perfect at all frequencies if the relation $L = Cr^2$ holds.

The reactance of carbon film type standards may be considered due entirely to capacitance between terminals. This may be held to about 1 $\mu\mu$f.

VARIATION WITH FREQUENCY. Resistance change with frequency is due principally to the presence of skin effect and of residual reactances. Skin effect may be reduced by using a small size of wire and reducing the inductance of the winding by suitable choice of winding type. For non-inductive windings, the increment due to skin effect may be taken as less than 0.1 per cent for No. 28 B.&S. gage wire at 1 megacycle, and less than 1 per cent for No. 30 wire at 10 megacycles. Where power considerations do not enter, sizes smaller than these are generally used in order to reduce the phase angle as already discussed, and skin effect is therefore usually negligible.

Reactance is a more serious cause of resistance variation with frequency. If a resistance has inductance in series with it, the effective resistance component of the combination considered as a parallel circuit will be increased to the value $r + (\omega^2L^2/r)$. If the resistance has capacitance in parallel with it, the effective resistance of the combination considered as a series circuit will be decreased to the value $r - \omega^2C^2r^3$. When both inductance and capacitance are present, the value of resistance is a function of both, and, if the inductance and capacitance are in such proportions as to give it a zero phase angle, its value will be increased to $r + (\omega^2 L^2/r)$. For high resistances at high frequencies, the dielectric loss of the associated capacitance affects the resistance value.

TYPES OF WINDINGS. The following windings are suitable for resistance standards having values above about 50 ohms, for which cases the reduction of both inductance and capacitance is important:

The Curtis winding made by winding one turn around a spool, then passing the wire through an axial slot and winding the next turn in the opposite direction. A flat slotted form or card is an improvement over a round spool.

The inductive winding on a thin flat card.

The woven type in which the warp consists of silk or cotton threads and the weft is the resistance wire.

The Ayrton-Perry type consisting of two parallel opposed windings either in a single layer, in which case they must cross at every turn, or one layer wound over the other. The single layer has a little more capacitance and a little less inductance than the two layers.

For resistance below 50 ohms where capacitance has less effect, windings having greater distributed capacitance such as the bifilar, and the reversed layer type may be used.

A well-made unit using No. 36 B.&S. gage wire may be wound to have a ratio of reactance to resistance, due to inductance, of about 0.1 at 1 megacycle, and, using No. 44 wire, to have a ratio of about 0.2 at 10 megacycles. If the capacitance is such as to give zero phase angle, which it may for windings of about 1000 ohms, the corresponding resistance increment will be about 1 and 4 per cent respectively. If the capacitance value is not such as to give zero phase angle, the effective shunt resistance will be different from the effective series resistance.

Table 2 gives representative values of inductance and capacitance for standards of several values wound with various sizes of wire.

Table 2. Representative Resistance Windings

Inductance and capacitance for various values and wire sizes

Resistance, ohms	Type of Winding	Insulation and Wire Size B.&S. Gage	Residuals		Net Residual	
			L, μh	C, $\mu\mu$f	L', μh	$C' = L'/r^2$, $\mu\mu$f
0.1	Bifilar, tape	0.004 in. × 0.125 in.	0.01	......	0.01	....
1	Bifilar, tape	0.004 in. × 0.063 in.	0.04	......	0.04	....
10	Bifilar	33 DSC	0.16	10	0.16	....
100	Bifilar	36 DSC	0.8	80	0	0
100	Curtis (card)	40 DSC	0.4	0.5	0.4	−40
100	Ayrton-Perry (2 layers)	40 BE	0.35	1	0.34	−34
100	Woven	40 BE	0.30	1	0.29	−29
1,000	Reversed layer	39 DSC	4.2	24	−20	20
1,000	Ayrton-Perry (2 layers)	44 BE	2.5	1	1.5	−1.5
1,000	Woven	44 DSC	2.0	0.5	1.5	−1.5
1,000	Curtis (card)	40 DSC	4.0	1.5	2.5	−2.5
10,000	Reversed layer	42 DSC	30	20	....	20
10,000	Ayrton-Perry (1 layer)	44 DSC	25	2	....	1.8
10,000	Woven	44 DSC	20	0.5	....	0.3

VARIABLE STANDARDS. Variable standards are used in precision measurements principally for covering the range between the smallest steps of the adjustable standards. They are usually wire wound and of low range. The simple series type has the objection that the contact resistance is usually an appreciable part of the total resistance of the standard. Preferred methods are to use the slide wire in a circuit as a potentiometer, in which case the contact resistance is not in the measuring circuit, or to use a shunted type of slide wire, in which case the resistance of the slide wire itself is higher than the range of the combination, thus reducing the effect of the contact resistance. Inductance constitutes the principal frequency limitation. The shunted type has lower inductance than the potentiometer type but does not have a linear scale unless used as a potentiometer. Compensation for inductance in single-turn slide wires may be made by substituting an equivalent length of copper wire for the resistance wire removed from the circuit. Compensation in the wound type is not generally made but can be effected in special cases by winding the slide wire non-inductively.

High-resistance variable standards are sometimes used in shunt connection as low-range conductance standards. These are commonly a composition type, particularly when used at high frequencies. They are not very stable standards.

ADJUSTABLE STANDARDS. Adjustable standards consist of a number of switch assemblies, usually 1 to 6, each containing resistance units arranged to give values in steps from 1 to 10. The switch and wiring add both inductance and capacitance. These may be reduced by reducing the switch size and the amount of dielectric material in it. These requirements are met very well by the wafer-type switch. The small size introduces some difficulties in obtaining satisfactory low and stable contact resistance, but the introduction of silver contacts reduces this difficulty. Their principal objection is their short life. They have considerable flexibility due to the availability of multiple decks. The usual arrangement is to connect 10 equal resistances in series and to short-circuit those not in use by means of the brush. The ground or low-potential side of the circuit should be connected to the brush to minimize the effect of capacitance. The inductance may be compensated for by using an additional 10 studs as shown in Fig. 1, to insert inductance equal to the inductance of the resistance and wiring which is removed, thus keeping the total circuit inductance constant for all switch settings at the cost of a slight increase of the zero inductance. This arrangement uses 2 decks of a wafer switch.

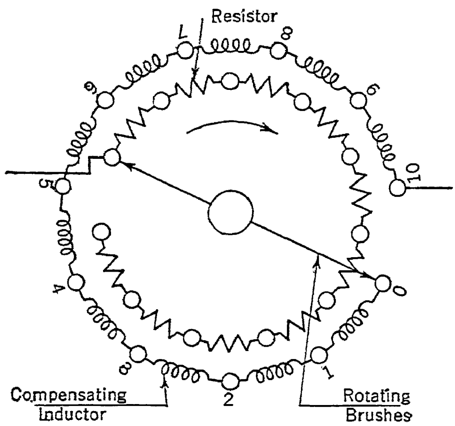


FIG. 1. Decade Resistance Standard, Compensated for Inductance

A method which reduces both capacitance and inductance to a minimum, and is therefore suitable for both high and low resistances, is shown in Fig. 2. A 20-stud switch is so arranged that the drum supporting the units and studs rotates, the brushes remaining stationary. Each unit in a decade has a resistance equal to the total value required for that particular setting. There is no switch wiring and there are no coils connected when not in use. The short-circuit connection for the zero setting may consist of a copper strap equal in inductance to the mean value of the units. The same arrangement with stationary studs and moving brushes is almost as satisfactory and allows the use of a 2-deck wafer switch. The principal disadvantage is the necessity for 10 different values for the resistance units.

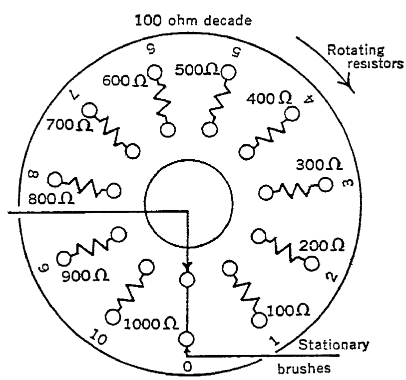


FIG. 2. Rotor Type Decade Resistance Standard

For high resistances there are advantages in arranging decades to read conductance. This means connecting them in parallel. Four units of value 1, 2, 3, 4 can be connected in various parallel combinations. The switching arrangement is the same as that for switching capacitors, shown in Fig. 5A.

9. CAPACITANCE STANDARDS

The principal requirements for a capacitance standard are a capacitance independent of time, humidity, temperature, voltage and frequency; a low power factor; and small size. The degree to which these requirements can be met depends principally on the choice of dielectric. All capacitors using solid or liquid dielectrics have losses when subjected to alternating voltages. The equivalent circuit of such a capacitor may be indicated as shown in Fig. 3 as a perfect capacitance C either with a resistance r_s in series with it, or with a conductance g in parallel with it. For these networks to be equivalent the two capacitances are not identical, but for the low power factors of standards, the difference is negligible.

Practically the only dielectrics used for capacitance standards are mica or air. Many other solid dielectrics are used for capacitors. None of them combines all the advantages of mica, although polystyrene is equal or superior to mica in practically all respects except temperature coefficient of capacitance.

Mica capacitors may be made of interleaved sheets of mica and foil such as copper, tin, or aluminum, or by depositing a film of metal, usually silver, directly on the mica. The last has the advantage of eliminating any air film between the dielectric and the metal, making the capacitance less dependent on the mechanical stability of the assembly.

Air capacitors require solid dielectric in their construction. Their performance depends considerably on their design as well as on the dielectric.

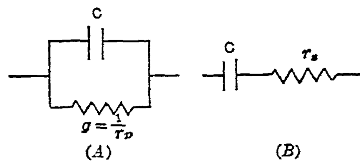


FIG. 3. Alternate Equivalents for an Imperfect Capacitor. *A*, parallel equivalent; *B*, series equivalent.

STABILITY WITH TIME AND HUMIDITY. All capacitors change slightly with age, owing to structural changes after assembly; good mica capacitors usually increase less than 0.05 per cent with age. Humidity has a serious effect unless the capacitors are protected. They are usually dried and hermetically sealed, either in molded plastic or in a can. In addition they may be vacuum-impregnated, usually with paraffin or Superla wax. Well-made air capacitors are less subject to these changes.

VARIATION WITH TEMPERATURE. The change in capacitance with temperature of mica capacitors is due partly to the dielectric material but is also affected considerably by the mechanical design. Unimpregnated capacitors usually have a positive linear change, which is due primarily to changes in the physical dimensions. In impregnated capacitors, additional changes occur owing to the effect of the wax, which has a negative temperature coefficient that increases with temperature. As a result, the temperature coefficient of the capacitor depends on the amount of wax and may be made extremely small over a limited temperature range. Standards can be obtained which vary less than 0.05 per cent from 10 deg cent to 35 deg cent. The shunt conductance of all mica capacitors increases with temperature. The amount varies considerably, but an average value may be taken as 2 per cent per deg cent. For air capacitors capacitance change is due partly to the insulating supports but principally to change in the plate area and spacing caused by changes in the dimensions of the parts. It may be reduced by the choice of materials of low temperature coefficient of expansion and by avoiding distortion due to unequal expansion of the parts.

VARIATION WITH FREQUENCY. Variations in capacitance with frequency of mica standards are caused by variations in the dielectric constant, and at high frequencies by the effect of inductance in the leads. The former causes a slight decrease in capacitance with increasing frequency, the change being approximately logarithmic with respect to frequency. The latter causes an increase in capacitance equal to ω^2LC^2. Lead inductance in a good standard may be held below 0.05 μh. Air capacitors change less, because of the smaller amount of dielectric material and because their comparatively low capacitance is affected by lead inductance only at higher frequencies.

An imperfect capacitor may be represented as shown in Fig. 3*A* or Fig. 3*B*. Metallic losses constitute a series resistance which is independent of frequency, except for skin effect. Leakage constitutes a shunt conductance independent of frequency. Dielectric loss is of such a nature that the loss per cycle is proportional to capacitance and approximately independent of frequency. In other words, this loss is proportional to ωC. If represented by Fig. 3*A*, g will be directly proportional to frequency, and if represented by Fig. 3*B*, r_s will be inversely proportional to frequency.

Table 3. Loss Variation with Frequency and Capacitance

Source of Loss	r	Q	P.F.	g
Series resistance r_s	r_s	$\frac{1}{r_s\omega C}$	$r_s\omega C$	$r_s\omega^2C^2$
Dielectric, D	$\frac{D}{\omega C}$	$\frac{1}{D}$	D	$D\omega C$
D-c leakage g_0	$\frac{g_0}{\omega^2C^2}$	$\frac{\omega C}{g_0}$	$\frac{g_0}{\omega C}$	g_0

In general, four different quantities are in common use for indicating the quality of a capacitor from the standpoint of loss. These are series resistance r, shunt conductance g, power factor which when small is equal to the dissipation factor ωCr or $g/\omega C$, and the reciprocal, which is commonly termed Q. Table 3 indicates how these quantities vary with frequency and capacitance.

In this table D is a fundamental property of the dielectric and is approximately independent of the frequency. For good mica it is as low as 0.0001 and decreases slightly with increasing frequency.

Low power factor or high Q in a standard is obtained by choice of mica of low dielectric loss, by keeping lead resistance low, and by eliminating moisture and impurities in assembly. Figure 4 shows curves of power factor and Q in terms of frequency for representative standards. The increase in Q with frequency is a characteristic of the dielectric. The decrease at high frequencies is due to effective series resistance. The larger the value of capacitance, the lower is the frequency at which the maximum occurs. Small capacitance values usually have greater dielectric loss because of the greater influence of the sealing material. They also have greater effective series resistance, owing, in part at least, to the smaller number of sheets in parallel.

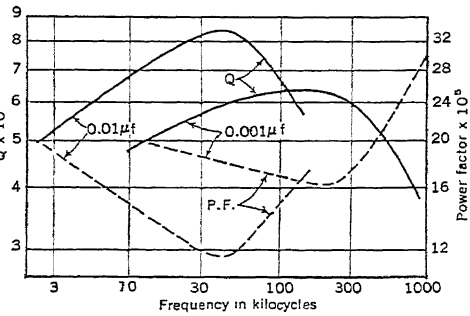


FIG. 4. Variation with Frequency of Q of Typical Mica Capacitor Standard

The Q of air standards is usually higher than that of mica on account of the reduced dielectric material and the reduced effect of series resistance due to the smaller capacitance.

VARIABLE STANDARDS. Variable standards are practically always air capacitors. They may be considered as two capacitances in parallel, one consisting of the minimum capacitance which includes the dielectric material, and the other consisting of the variable part. The former may be considered as a fixed capacitance shunted by a conductance representing the dielectric loss, which is not a function of setting. The latter may be considered as a loss-free variable capacitance in parallel with the former. They may be used, therefore, in measuring circuits, to obtain loss-free changes in capacitance. On the other hand, considering the standard as a whole, the power factor increases with decrease in setting and may be greater than that of a mica capacitor at minimum setting unless a very low loss dielectric is used.

ADJUSTABLE STANDARDS. These are usually in decade form, but, owing to the cost of good standards, only 4 units are generally used, the switch connecting them in the parallel combinations necessary to obtain steps from 1 to 10. Practically all such switches reduce to the principle of a brush for each unit contacting successively 10 studs which are wired so as to connect the unit in circuit at the positions required to give the desired combinations. This may be most readily done with minimum wiring by using cams. The method is indicated by the developed cam arrangement shown in Fig. 5*A*. By using 2 brushes on different diameters a single cam may be cut to connect 2 units. Wafer switches can be used requiring only 2 cams and 5 brushes mounted on a single deck to perform the whole switching sequence. A similar switch may be mounted on the same shaft with cams cut inversely as shown in Fig. 5*B* but parallel-connected to insert capacitances in place of those removed to compensate for the capacitance of the switch or for that of the units in excess of nominal values.

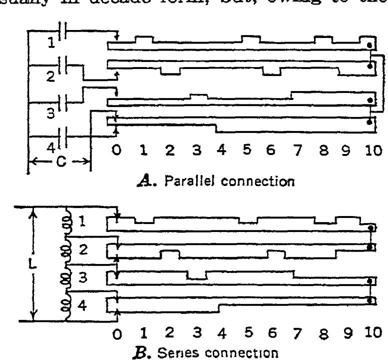


FIG. 5. Cam Type of Decade Switch for Series or Parallel Connection of 4 Standards. *A*, parallel connection; *B*, series connection.

10. INDUCTANCE STANDARDS

The requirements for inductance standards are similar to those for capacitance standards, namely, high stability of inductance under all operating conditions, low power factor or high Q, and small size. Coils cannot be built to meet these requirements as well as capacitors, and for this reason they are not used to the same extent as standards. For precise standards no magnetic material is sufficiently stable, and air-core coils are almost always used. They are wound either as toroids or as solenoids. The toroids have less external field but are larger for the same performance. External field is objectionable since it not only causes errors in measurement due to coupling to the rest of the circuit but the inductance of the standard itself is affected also. Shielding is generally used consisting of magnetic material at low frequencies and low-resistance material such as copper at high frequencies.

STABILITY WITH TIME AND HUMIDITY. Coils may be constructed to have negligible change in inductance and resistance with time. They are usually sealed and in addition may be impregnated to eliminate humidity effects.

VARIATION WITH TEMPERATURE. Change in inductance is determined principally by mechanical changes in the form and in the impregnating material when present. By choice of materials of low expansion these changes can be held to as low as 0.001 per cent per deg cent. A single-layer solenoid wound under tension on a ceramic form may be made with a temperature coefficient of less than 2 parts per million per deg cent. Change of resistance is due principally to change in the resistivity of the copper winding and to change in eddy-current loss which in turn is due to the change in resistivity. These changes are of opposite sign. The net change is a function of frequency but is usually positive.

VARIATION WITH FREQUENCY. Variations of inductance and effective resistance with frequency are due principally to eddy currents and distributed capacitance. The former results in an increment of resistance which is proportional to the square of the frequency. It can be reduced by using stranded wire and single-layer windings. The effect on the inductance is usually negligible compared with other effects. Distributed capacitance has the effect of increasing both inductance and effective resistance. It is a minimum for single-layer coils. Where this type is impractical, banked or sectionalized windings may be used. Capacitance to ground can be reduced appreciably only by reducing the size of the coil, which is at the cost of Q. For frequencies well below the resonant frequency of the coil, the increment in inductance due to capacitance C is ω^2L^2C, and the increment in the effective resistance is $2\omega^2LCr$. The dielectric conductance g associated with the distributed capacitance may also be appreciable. It increases the resistance of the coil by $g\omega^2L^2$. If the coil has no appreciable resistance or inductance variation with frequency, then the ratio of reactance to resistance or Q is proportional to the frequency. On the other hand the loss due only to capacitance and eddy currents results in a Q inversely proportional to the frequency. For the actual case, the Q usually increases to a maximum and then decreases. For low frequencies a high Q can be obtained only by means of a large size. This is the principal disadvantage of air-core standards. It is most evident when they must be used for measuring magnetic-core coils. As the frequency increases the comparison becomes more favorable to the air-core coils on account of the increase in core loss and decrease in effective permeability of magnetic core coils. Above 100 kc this disadvantage is negligible, and air-core coils can be built of small size to have values of Q as high as 100 to 200.

INDUCTOMETERS. All inductometers have an appreciable stray magnetic field which may cause error due to magnetic coupling to other parts of the circuit. They have characteristics similar to fixed standards but usually have relatively greater capacitance and greater resistance increment. The inductance and resistance increments are a function of the setting as well as of frequency.

One type consists of two coils arranged so that the position of the field of one may be varied so as to change their mutual inductance. They may be used as variable mutual inductors or, by connecting the two windings in series, as variable self-inductors. Another type consists of a single-layer rotating solenoid with spaced bare wire turns and a fixed brush riding on the turns. The former type has the advantage of no sliding contacts, but the latter has better performance in other respects.

ADJUSTABLE STANDARDS. Coils may be assembled to form adjustable standards similar to resistance and capacitance standards. Because of their large size, poorer characteristics, and difficulty in compensating for capacitances of switch and wiring with series connection, they are not used extensively. A typical method of connecting 4 units to form a decade is shown in Fig. 5B. Additional decks with cams cut inversely may be used to insert compensating resistances in place of inductances removed.

11. MEASUREMENT OF RESISTANCE

The method used almost exclusively for measuring circuit constants is the bridge method in which the unknown quantity is compared with a standard of known value. For the higher frequencies, say, above 10 megacycles, the errors caused by the presence of even very small stray inductance or capacitance limit the range and flexibility of the bridge method, and consequently at these frequencies other methods which involve less apparatus and simpler circuits are also used.

Resistance standards can be made so that their effective resistance does not differ from the d-c value appreciably compared with the errors of measurement up to the maximum frequencies used. Accordingly, the measurement of resistance is usually made by direct comparison with a standard on a bridge. The simplest type is the equal ratio-arm bridge. A modified equal ratio-arm bridge useful for the a-c measurement of both resistance and phase angle is shown in Fig. 6. It is adaptable only to the measurement of resistors having small phase angles, but, with a given air capacitor C_b of small range across one ratio arm, it will measure the reactance of resistors of any value over a wide frequency range. The equations of balance for this bridge for resistors of small phase angle are

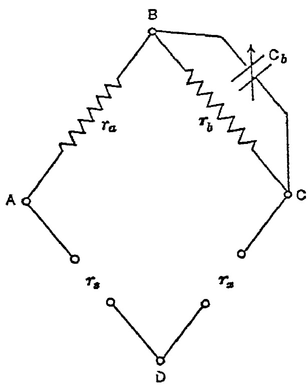


FIG. 6. Bridge Circuit for Measurement of Resistance and Phase Angle of Resistors

$$C_x = \frac{C_b r_b}{r_s} \quad \text{or} \quad L_x = -C_b r_b r_s$$

and

$$r_x = r_s$$

where r_s is the resistance of the standard.
r_x, C_x, and L_x are the values for the unknown.

Both positive and negative reactance can be measured by either transposing the standard and unknown in the bridge or by transferring the capacitance C_b from BC to AB. If the standard has appreciable reactance, correction for it is necessary. For maximum accuracy the bridge should be shielded. A method of shielding similar to that shown in Fig. 8 is satisfactory.

At very high frequencies a direct substitution in a resonant circuit as described under "Measurement of Inductance" may be used, or a voltmeter-ammeter method, with a vacuum tube voltmeter and a thermocouple.

Calibration of standards for phase angle may be made by substitution of one resistance by another having the same physical configuration but a different resistance value, thus obtaining a change in resistance without changing the associated inductance or capacitance. For low resistances this may be done by making identical standards of copper and of various resistance alloys. For high resistances, carbon of different thickness may be deposited on identical forms.

The measurement of effective resistance associated with inductance and capacitance is considered under those headings.

12. MEASUREMENT OF CAPACITANCE AND CONDUCTANCE

TYPES OF CAPACITANCE. In practice, a capacitor does not usually consist of a single capacitance between two terminals. There is, in addition, capacitance to ground or to other objects. The capacitor is usually equivalent to a three-terminal network of three capacitances even when reduced to its simplest form. As a result, different values of capacitance will be obtained, depending on the conditions of measurement. Referring to Fig. 7, C_1 is the capacitance to ground from A, and C_2 is the capacitance to ground from B. The direct capacitance between A and B is C; the mutual capacitance is $C + C_1C_2/(C_1 + C_2)$; and the grounded capacitance is $C + C_1$ or $C + C_2$, depending on whether B or A is grounded. Any or all of these capacitances may require measurement.

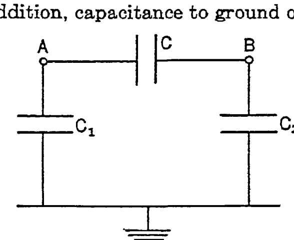


FIG. 7. Simplified Network of Capacitor Which Has Capacitance to Ground

In actual use, the capacitor is generally connected either from one side of the system to ground, in which event it is a grounded capacitance which is effective, or directly across the system which is balanced to ground, in which event it is the mutual capacitance that is effective. Direct capacitance is required, in general, only for special purposes or where the capacitor is so connected in service that the three direct capacitances are connected across different parts of the circuit and must be known individually.

BRIDGE METHODS OF MEASUREMENT. Owing to the satisfactory characteristics of capacitance standards, the measurement of capacitance is usually made by direct comparison with them. The simplest and most accurate method is the equal ratio-arm bridge. The advantages of this type of bridge over all other types are that the equality of the ratio arms may be tested in the bridge itself by means of simple reversal, and residual and lead impedances may be more readily balanced because of the symmetrical arrangement. The ratio arms are usually resistances. Capacitances have advantages for high voltages, and inductances, particularly when closely coupled, for high currents. For precise work, especially at the higher frequencies, some method of taking care of undesired capacitances between arms and from the arms to ground is necessary, the best method being to shield the bridge. This shielding should include at least one transformer. The shield should be complete enough to reduce the direct capacitance between windings below 1 $\mu\mu f$. In special cases it should be much better than this.

Capacitance. Figure 8 shows a shielded equal ratio-arm bridge which is satisfactory for the measurement of capacitance by direct comparison with a standard. This bridge will measure mutual, grounded, and direct capacitance. The requirement that must be met in order to measure mutual capacitance is that the bridge corners C and D be balanced to ground. Since there is no capacitance to ground from B, and since A and C are at the same potential when the bridge is balanced, this requirement is met when the total capacitance from A and C to ground equals the capacitance from D to ground. If at the same time these capacitances are proportioned to keep the bridge in balance, then the capacitances from A and C to ground will be equal, each being half the capacitance from D to ground. This adjustment may be made by means of auxiliary air capacitors. Grounded measurements are made by grounding the D corner by switch K_1. If the bridge is limited to grounded measurements, D may be permanently grounded, and the double shielding of the transformer T_1 and the ratio arms is unnecessary.

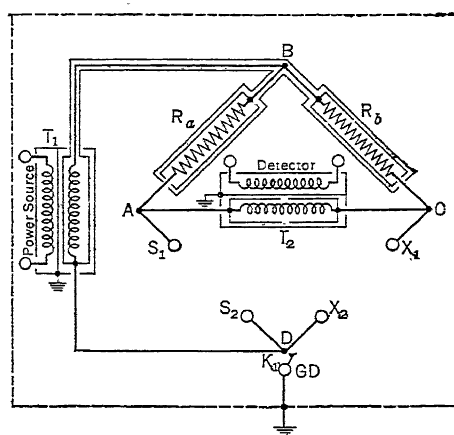


FIG. 8. Completely Shielded Equal Ratio-arm Bridge Suitable for Balanced-to-ground or Grounded Measurements

Conductance. The loss in a capacitor may be measured as a series resistance or as a shunt conductance. Using the bridge of Fig. 8 a standard resistance may be connected either in series or in parallel with the standard capacitor to balance this loss. For wide ranges of capacitance and power factor, however, the range of resistance required in either case is objectionable. The series method may require excessively small resistances, and the shunt method may require excessively high resistances. A modification which permits of a practical resistance range is to shunt both arms by a resistance of fairly high value and to reduce the resistance across the standard to balance any loss in the capacitor under test. Satisfactory values are 10,000 ohms for the fixed resistance across CD and a variable resistance range from 0.01 ohm to 10,000 ohms across AD. This allows the measurement of conductances from 0.0001 micromho up. If r_0 is the reading of r_{AD} in ohms for the zero balance and r_1 is the reading with the capacitor in circuit, then

$$g_x = \frac{r_0 - r_1}{r_0 r_1} \quad \text{mhos}$$

If r_0 is approximately 10,000 ohms, and the conductance is below 1 micromho, this reduces to

$$g_x = \frac{r_0 - r_1}{100} \quad \text{micromhos}$$

The bridge then becomes practically direct reading for conductance. This approximation holds for conductance values encountered in all good standards at moderate frequencies. These expressions are actually the difference in the loss between the standard and the capacitor under test. If the standard has conductance, it must be added to the difference obtained.

In order to calibrate standards for conductance, it is necessary to have a primary standard of zero or known conductance. This is usually obtained by means of an air capacitor specially designed to give a direct capacitance having no loss. If an air capacitor is built so that each set of plates is mounted by insulating supports on a third metal conductor or shield, we then have a system of three capacitors having three direct capacitances such as shown in Fig. 7, in which the dielectric loss is wholly in the direct capacitances from the plates to the support, the direct capacitance between the plates including no dielectric loss. This method, or the use of a variable air capacitor with conductance independent of setting, is the common method for obtaining standards of zero conductance.

Accuracy. The accuracy of a bridge of this type depends on the accuracy with which the capacitance standards are known. An overall accuracy of 0.1 to 0.25 per cent for capacitance is possible for frequencies up to 1 megacycle. The conductance accuracy stated in terms of power factor is in the order of $\pm(1$ per cent $+ 0.0001)$.

DIRECT CAPACITANCE. Direct capacitance may also be measured on the bridge of Fig. 8 as follows: Ground the point C of the bridge and all conductors of the apparatus under test except the two terminals between which the direct capacitance is desired. Connect one of these to C and one to D, and obtain a bridge balance. Then transfer the terminal connected at C to A, and rebalance the bridge. A little consideration will show that the only capacitance transferred in this operation is the direct capacitance required, and the difference between the balances is therefore equal to twice this capacitance.

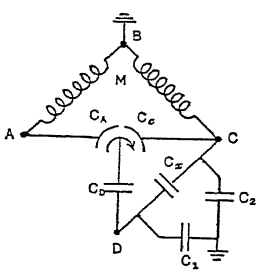


FIG. 9. Bridge for Measurement of Direct Capacitance

A second method which is satisfactory where all the capacitances involved are small, such as the interelectrode capacitances of a vacuum tube, is to connect the capacitance to be evaluated across CD, all other terminals including ground being connected to B. The capacitance thus placed across BC will affect the accuracy of conductance but not of capacitance measurement, and the direct capacitance may be measured directly.

The modified circuit shown in Fig. 9 has two advantages for the measurement of direct capacitance. By using a pair of closely coupled inductors for ratio arms, that is, parallel windings on a common core, the effect of the capacitance across BC will be very much reduced. If, in addition, the capacitance standard is such that capacitance is transferred from CD to AD, keeping the total value $C_A + C_C$ constant, and a fixed capacitance C_D is connected in series in the D lead, it may be shown that the bridge unbalance corresponding to an unbalance $C_A - C_C$ is equal to $\dfrac{(C_A - C_C)\, C_D}{C_A + C_C + C_D}$. If C_D is made small compared with $C_A + C_C + C_D$, very small values of direct capacitance, as low as 0.0001 $\mu\mu$f, may be measured.

A third method is by means of the Wagner ground.

The Wagner Ground. A common modification of the simple equal ratio-arm bridge and of other bridges involves the Wagner ground. This is a simple method of eliminating the effect of stray admittances, compared with complete shielding of the bridge. Referring to Fig. 8, but with power source and detector interchanged, it consists of a potentiometer connected across the input corners AC of the bridge, the adjustable contact being grounded and adjusted to bring the D corner of the bridge to ground potential. In some bridges, complex impedances are required instead of resistances for this balance. It has two disadvantages. First, it requires for any measurement two balances, namely, the adjustment of the Wagner ground and of the bridge itself, and these balances are not independent. This makes the balance procedure relatively slow and complicated. Second, the quantity measured is the direct capacitance. When the mutual or grounded capacitance is desired, it must be computed from the measured values of the individual direct capacitances. These limitations considerably restrict its usefulness.

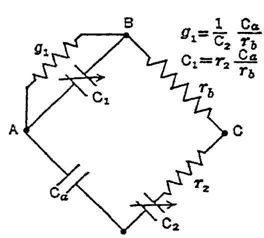


FIG. 10. Schering Bridge for Measurement of Capacitive Impedances

THE SCHERING BRIDGE. At high frequencies, adjustable resistance standards have serious limitations. Above about 1 megacycle, the Schering bridge is used to avoid them. In this bridge a fixed capacitance and a fixed resistance are in diagonally opposite arms, and the other arms may be arranged to allow both capacitance and conductance to be balanced by variable capacitors. The circuit and balance equations are given in Fig. 10. If the bridge is balanced and then the unknown is connected across C_1 and the bridge rebalanced by means of C_1 and C_2, the change in C_1 equals the capacitance of the unknown and the change in C_2 is inversely proportional to its conductance. If series components are desired the unknown may be connected either in series or in parallel with C_2. Then C_2

balances for capacitance and C_1 for series resistance but the bridge will not be direct reading for both components.

OTHER METHODS. Bridge methods have the advantage of a null balance. They do not have the advantage of both grounded input and grounded output unless transformers are used. Other methods are available, particularly at high frequencies, which have both these advantages. They consist essentially of two unbalanced transmission networks connected in parallel, the required relation between the arms for zero output being that their transfer constants be equal in magnitude but 180° out of phase (see Section 6). Commercial measuring circuits of this type are available, known as shunted T and twin T circuits.

The most popular method other than null methods is the simple tuned circuit, the condition sought being a maximum or minimum of either voltage or current with variation of inductance, capacitance, or frequency. A commercial circuit using this method known as a Q meter is described in article 13.

Practically all these special circuits substitute a capacitance standard for the unknown in measuring capacitance, the difference between them lying in the method of measuring the loss. Capacitance is seldom measured in terms of other quantities owing to the superiority of capacitance standards over all others.

13. MEASUREMENT OF INDUCTANCE AND EFFECTIVE RESISTANCE

As for capacitance, the simplest type of measurement is direct comparison, and the bridge of Fig. 8 is capable of the same accuracy for inductance measurements as for capacitance measurements, limited only by the standards. However, inductance standards are inherently less satisfactory than capacitance or resistance standards, particularly for wide inductance and frequency ranges. Accordingly, other less symmetrical bridges are common in order to take advantage of standards of capacitance and resistance. Of these the bridges most used are the resonance method with the equal ratio-arm bridge, the Owen bridge, the Maxwell bridge, and the Schering bridge.

COMPARISON METHOD. This measurement is made by comparing an inductance in the CD arm with an adjustable standard in the AD arm of an equal ratio-arm bridge such as shown in Fig. 8. In order to balance the effective resistance component and to measure it when required, it is necessary to add a series resistance in either the CD arm or the AD arm, depending on the relative resistance values of the unknown impedance and of the standard. The usual procedure is to connect a standard resistance by means of a switch at D to throw the resistance into either arm as required. The bridge permits balanced to ground and grounded measurements, as in capacitance measurements. When adjustable standard inductances are used, and when the series resistance required is large, the shielding of the standards becomes cumbersome and the bridge is commonly used grounded.

The inductance of the unknown is equal to the inductance of the standard at balance. The effective resistance is given by $r_x = r_L \pm r_s$ where r_L is the effective resistance of the standard inductance and r_s is the setting of the standard resistance, the sign depending on the position of the switch.

The accuracy of such a circuit depends on the frequency and on the accuracy of calibration of the inductance standards. An accuracy of about 0.25 per cent for inductance and 2 per cent for resistance is possible for frequencies as high as a megacycle.

Wagner Ground. This can be used with the above method to avoid shielding. However, the limitations of this method as outlined in article 12 apply equally to inductance measurements.

RESONANCE METHOD. Inductance can be compared with capacitance and frequency either by series or parallel resonance. The series method has the advantage of giving directly the series resistance and inductance of the coil which are the values usually desired. It is also inherently a low-impedance circuit, and this is often an advantage where the voltage available from the power source is limited. The parallel method gives the equivalent parallel values, which usually require subsequent transformation. It is inherently a high-impedance circuit.

The principle on which resonance measurements are based is the adjustment of the capacitor until the tuned circuit has zero or infinite reactance; that is, it is equivalent to a pure resistance. The measurement is usually made on an equal ratio-arm bridge, but any bridge that will determine when the impedance of the circuit is a pure resistance and that will measure the resistance is suitable. The principal objections to the method are that it is not direct reading and the accuracy is dependent on the frequency of the source.

Series Resonance. For the series measurement by the equal ratio-arm bridge of Fig. 8, an adjustable resistance is connected in the AD arm and the capacitance in series with the unknown in the CD arm. Since the unknown forms only part of the impedance in the CD arm, balanced-to-ground measurements are impractical and this measurement is usually made with D grounded. The capacitance is connected from C to the unknown in order that one terminal of the unknown may be connected to ground. The balance is obtained by adjusting the standard resistance and capacitance, and at balance the following relation holds:

$$\omega^2 L_x C_s = 1 \quad \text{or} \quad L_x = \frac{1}{\omega^2 C_s}$$

r_x is equal to r_s less the equivalent series resistance of C_s. If the conductance of C_s is used, as it generally is, since when multiple standards are used their conductances add directly, then

$$r_x = r_s - \frac{g_s}{\omega^2 C_s^2} = r_s - \frac{L_x g_s}{C_s}$$

where r_x, L_x are the values of the unknown.
g_s is the conductance of the standard C_s.

Parallel Resonance. For this measurement the capacitance and the unknown inductance are connected in parallel in the CD arm of the bridge. Either balanced-to-ground or grounded measurement may be made. Since the resistance required for the balance may be very high, an arrangement similar to that for a capacitance bridge is customary; that is, a fixed resistance, usually 10,000 ohms, is connected across CD and the loss is read as conductance by means of a variable resistance across the AD arm of the bridge. The balance is then obtained in the same way as for series resonance. Then

$$L_p = \frac{1}{\omega^2 C_s} \quad \text{and} \quad g_p = \frac{r_0 - r_1}{r_0 r_1} - g_s$$

where g_s is the conductance of the condenser C_s.
r_0 and r_1 are the open-circuit and final readings of r_s.
L_p is the parallel inductance of the unknown impedance.
g_p is its conductance.

The series equivalents may be obtained by transformation.

Accuracy. The accuracy of resonance bridges depends on the accuracy of both the capacitance standards and the frequency. Accuracies as high as 0.1 per cent are possible without extreme precautions. They are probably the most accurate circuits for measuring effective resistance, accuracies of the order of 2 per cent being readily obtainable, even for high-Q coils.

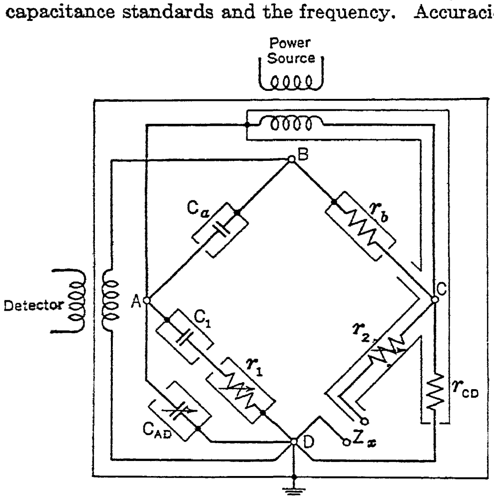


Fig. 11. Shielded Owen Bridge for Measurement of Inductance and Effective Resistance

THE OWEN BRIDGE. This bridge in common with the Maxwell bridge has the advantage that inductance is measured in terms of capacitance and resistance. It is not frequency sensitive and can be made direct reading for both L and r. A shielded circuit for the Owen bridge is shown in Fig. 11. The equations of this circuit at balance are

$$L = C_a r_b r_1 \quad \text{and} \quad r_2 = \frac{C_a r_b}{C_1}$$

where r_2 includes the effective resistance of the unknown. Two methods of operation are possible. Both C_1 and r_1 may be adjustable, and then the inductance of the unknown is proportional to r_1 and the total effective resistance in the CD arm is proportional to $1/C_1$. It is usually more convenient to have C_1 fixed or adjustable in a few steps and to have r_2 adjustable. Then by taking a short-circuit reading r_0 for r_2 it follows that $L = C_a r_b r_1$ and $r_x = r_0 - r_2$. Since the unknown is not connected directly across the arm CD, balanced-to-ground measurements are impractical. Since the resistances are capable of adjustment to their nominal values with a high degree of precision, the bridge is direct reading for both inductance and resistance without the need for calibration. Also, the range of inductance may be made very wide since r_1 may be designed to have as many as six or seven dials. The bridge is not so suitable for the measurement of effective resistance on account of residual reactances in the various bridge arms and difficulties in shielding always encountered in arms having series impedances.

The shielding is made only partially complete in the interest of simplicity. The resistance r_{CD} is used to compensate the shield capacitance across AD, a resistance being required because of the 90° phase relation of the ratio arms. With this circuit inductance accuracies as high as 0.1 per cent are usual at audio frequencies. It is not as satisfactory at higher frequencies as the following bridges.

THE MAXWELL BRIDGE. In this bridge fixed resistors are used in diagonally opposite arms. The inductance and effective series resistance of the unknown are then balanced by capacitance and conductance respectively in the standard arm. The circuit and balance equations are shown in Fig. 12. If r_a and r_b are not pure resistances, the balance equations still hold if their phase angles are equal but of opposite sign. If g_s is designed to read directly in conductance (see article 8) the bridge may be made direct reading for both L and r. This feature makes the bridge as attractive as the Owen bridge. In addition, since it has no series connections in the arms, stray admittances may be compensated more readily and the bridge is satisfactory at higher frequencies.

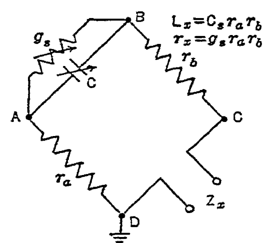


FIG. 12. Maxwell Bridge for Measurement of Inductance and Effective Resistance

THE SCHERING BRIDGE. This is described in article 12. It is essentially a capacitance bridge but will measure inductance as a negative capacitance and is particularly adapted to high-frequency measurements of low-Q impedances which may have either positive or negative reactance.

METERING. The values of inductance and resistance of magnetic-core coils are usually a function of the saturation. It is therefore desirable to know the current through, or the voltage across, the coil when measured. In any so-called ratio-arm bridge, where two adjacent arms are invariable, the current through the unknown, or the voltage across it, will have a definite relation to the total bridge current or voltage, determined only by the ratio arms and the choice of input and output corners. Thus in Fig. 8, which uses a current connection, the current input divides at balance in proportion to the admittances of the ratio arms, and a meter in the input circuit can be calibrated to read directly the current through the unknown. In Fig. 11, which shows a voltage connection, the input voltage divides at balance in proportion to the impedances of C_a and r_b, and a meter across the input can be calibrated to read directly the voltage across the unknown. Here the calibration will be a function of frequency but is independent of the value of the unknown.

Bridges of the so-called product-arm type, having the fixed impedances diagonally opposite, such as the Maxwell and the Schering bridges, do not have this feature, and the metering is usually done by connecting a vacuum-tube voltmeter directly across the unknown.

SUPERIMPOSED MEASUREMENTS. Measurement of inductance is sometimes required for magnetic-core coils, with direct current flowing through the winding. Such measurements may be made on an equal ratio-arm bridge such as that of Fig. 8, applying the direct current across the BD corners and separating the direct current from the alternating current where necessary by stopping condensers and a choke coil. A more convenient bridge is the Owen bridge of Fig. 11. If direct current is applied across BD and alternating current across AC, the only additional apparatus required is a stopping condenser in the detector circuit. The Maxwell bridge is also well adapted to making these measurements.

OTHER METHODS. At very high frequencies where extreme circuit simplicity is desirable, the simple tuned circuit is commonly used. The method of measurement is as follows.

If a voltage is applied to an inductance in series with a capacitance as shown in Fig. 13 and either L, C, or ω is varied to give maximum voltage across C, then

$$\omega L = \frac{1}{\omega C}$$

and

$$\frac{E_2}{E_1} = \frac{\omega L}{r} = \frac{1}{\omega C r}$$

where E_1 is the voltage across the tuned circuit.

E_2 is the voltage across L or C.

r is total resistance of the circuit.

Thus, if the values of E_1, E_2, and ω are measured, the inductance and resistance of a coil can be determined if the capacitance and resistance of the capacitor are known, and vice versa. A refinement of this circuit consists of adjusting the input voltage to a definite value either directly by a voltmeter or by adjusting the current into the known resistance r_c. Then the voltmeter across C may be calibrated in terms of Q. If the capacitor C has negligible loss, this will be the Q of the coil. Commercial measuring sets called Q-meters are available with self-contained oscillator, low-loss variable-capacitance standard, and vacuum-tube voltmeter, which are direct reading for Q. They are satisfactory up to 200 megacycles or higher.

There are a number of errors in the Q reading that may be corrected for when known. The resistance r of the circuit includes the equivalent series resistance of the capacitor, which includes its dielectric loss and the admittance of V_2. The frequency response of the voltmeters may not be flat. If the input current is controlled, E_1 is affected by change in impedance of r_c due to reactance or skin effect, and by the shunting effect of r. These errors vary independently with Q, X, and ω so much that no average figure for accuracy applies. For values of Q about 100, and values of X about 100 ohms, accuracies of 5 per cent can be expected up to 25 megacycles.

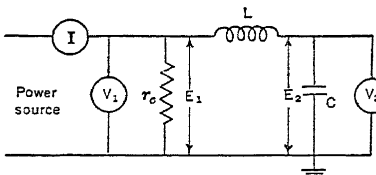


FIG. 13. Q Meter Circuit

MUTUAL INDUCTANCE. Mutual inductance between two coils can be determined by measuring the self-inductance of the two windings by any suitable method, with the windings connected first series aiding and then series opposing. The difference between the two values is four times the mutual inductance. The ground conditions are usually different for the two measurements from the actual operating conditions. This may be a source of error, particularly where the coupling is low. The ratio of secondary voltage to primary current may be determined directly by thermocouple and vacuum-tube voltmeter.

14. SIGNAL GENERATORS AND DETECTORS

The following discussion is limited to the special requirements which apply to use with measuring circuits such as those described. More complete information may be found in articles 36 to 40.

SIGNAL GENERATORS. Many single-frequency generators are available in the audio-frequency range provided the requirements are not severe. However, the tuning fork operated by a microphone or vacuum tube, and the rotating generator, represent practically the only types, other than vacuum-tube oscillators, with satisfactory characteristics for precise work. Where a range of frequencies is necessary and particularly for frequencies above the audio range, vacuum-tube oscillators are used almost exclusively.

VACUUM-TUBE OSCILLATORS. The principal requirements for an oscillator for measurement purposes are adequate output level, low level of harmonics and other spurious frequencies, and high stability of frequency and output level with respect to time and temperature. The emphasis on these requirements depends on the type of measurement. The output level should be high enough to insure that the input level to the detector is above thermal noise for the most precise balance. An output of 0.1 watt is usually adequate. Level stability requirements are more lenient for null measurements. Harmonic requirements are more lenient for null measurements which are not frequency dependent. In general, when untuned detectors are used, harmonics should be held below 3 per cent of the fundamental, and for null resonant measurements, below 1 per cent. Harmonics may be suppressed by filtering external to the oscillator, either before or after the measuring circuit. The former is preferable in so far as it prevents production of false signal by modulation in the measuring circuit, but it places severe modulation requirements on the filter.

Oscillators are in general of three types, according to the type of frequency-selecting network used, namely, crystal, LC, and rC oscillators. The crystal oscillator is the most stable and is the preferred type for fixed frequency applications over the frequency range for which crystals are applicable, about 10 kc to 10 mc. The LC oscillator is the most versatile and can be used over the whole frequency range up to the maximum frequency at which lumped constants are practical. At low frequencies, the necessity of using coils of large physical size to obtain a high Q makes it cumbersome, and the rC oscillator is preferred. This oscillator has the advantages that the components are small even at very low frequencies, and it may be made direct reading more readily. It is suitable for frequencies lower than 1 cycle up to about 100 kc, where, in spite of lower stability, it competes with the LC oscillator because of the direct-reading feature.

HETERODYNE OSCILLATORS. These oscillators, having an output frequency which is the difference between the frequencies of a fixed and a variable oscillator, can be made to cover a wide frequency range with a single continuous control. This has two advantages: it allows them to be made direct reading over wider ranges than the conventional *LC* oscillator, and the broad continuous range makes them suitable for sweep frequency measurements required in recording, and in cathode-ray visual indication, of frequency characteristics.

DETECTORS. Detectors, as distinct from actual measuring instruments, are limited to the detection of null balances and equality of output from different circuits or from different arrangements of the same circuit.

The principal requirements are adequate sensitivity for the accuracy required, sufficient discrimination against harmonics, and stability of gain with time and temperature. Sensitivity can usually be obtained by amplification. The limit is determined by thermal noise, which depends on the band width. This limit is approximately 10^{-20} watt per cycle band width. For instance, using a detector with a 1000-cycle band width working out of a bridge of 1000 ohms impedance, minimum signal to exceed thermal noise will be 10^{-17} watt or 0.1 μvolt.

The harmonic suppression required depends on the oscillator harmonics, the method of measurement, and the characteristics of the unknown. Gain stability requirements are least severe for null measurements. They are most severe in adjusting two outputs to equality using some form of suppression.

The telephone receiver with or without preceding amplification is the simplest and most sensitive detector within the audio-frequency range. Owing to the frequency characteristic of the ear and of the receiver, considerable discrimination against harmonics can be obtained in the frequency range where they are most sensitive. At higher frequencies, the heterodyne type of detector giving an audio output may be used with the receiver. It has the advantage of giving considerable discrimination. In all detectors care must be taken that the harmonics do not overload the input sufficiently to cause a false signal due to modulation.

In addition to the telephone receiver, a rectifier, such as copper oxide or a vacuum tube, may be used with a d-c instrument or cathode-ray-tube indicator. These require greater amplification and more discrimination.

BIBLIOGRAPHY

Campbell, A., and Childs, E. C., *The Measurement of Inductance, Capacitance, and Frequency*, Macmillan & Co., London, 1935.
Edgecumbe, K. W. E., and Ockenden, F. E. J., *Industrial Electrical Measuring Instruments*, Pitman, London, 1933.
Terman, F. E., *Radio Engineer's Handbook*, McGraw-Hill, New York, 1943.
Chaffee, E. L., *Theory of Thermionic Vacuum Tubes*, McGraw-Hill, New York, 1933.
Moullin, E. B., *Radio Frequency Measurements*, Griffin, London, 1926.
Hund, A., *High Frequency Measurements*, McGraw-Hill, New York, 1933.
Hague, B., *Alternating Current Bridge Methods*, Pitman, London, 1923.
Glazebrook, Sir R., *Dictionary of Applied Physics*, Macmillan & Co., London, 1923.
Hartshorne, L., *Radio Frequency Measurements by Bridge and Resonance Methods*, Chapman and Hall, London, 1941.
Henney, K., *The Radio Engineering Handbook*, McGraw-Hill, New York, 1941.
Rider, Jr., *Vacuum Tube Voltmeters*, J. F. Rider, New York, 1941.
Bur. Standards Circ. 74, Radio Instruments and Measurements.
Grondahl, L. O., The Copper-cuprous-oxide Rectifier and Photoelectric Cell, *Rev. Modern Physics*, Vol. 5, 141 (1923), including extensive bibliography on copper oxide rectifiers.
Aiken, C. B., Theory of Diode Voltmeters, *Proc. I.R.E.*, Vol. 26, 859 (1938).
Miller, J. H., Thermocouple Ammeters for Ultra High Frequencies, *Proc. I.R.E.*, Vol. 24, 1567 (1936).
Behr, L., and Tarpley, R. E., Design of Resistors for High Frequency Measurements, *Proc. I.R.E.*, Vol. 20, 1101 (1932).
Campbell, G. A., The Shielded Balance, *Elec. World*, Vol. 43, 647 (1904).
Ferguson, J. G., Shielding in High Frequency Measurements, *Trans. A.I.E.E.*, Vol. 48, 1286 (1929).
Campbell, G. A., Measurement of Direct Capacities, *Bell Sys. Tech. J.*, Vol. 1, 18 (1922).
Christopher, A. J., and Kater, J. A., Mica Capacitors for Carrier Telephone Systems, *Trans. A.I.E.E.*, Vol. 65, 670 (1946).
Cone, D. I., Bridge Methods for Alternating Current Measurements, *Trans. A.I.E.E.*, Vol. 39, 1743 (1920).
Behr, L., and Williams, A. J., The Campbell-Shackelton Shielded Ratio Box, *Proc. I.R.E.*, Vol. 20, 969 (1932).
Wagner, K. W., Zur Messung dielektrischer Verluste mit der Wechselstrombrücke, *Elekt. Zeits.*, 32 Jahrgang, pp. 1001–1002.
Kupfmuller, Von K., Über eine technische Hochfrequenz Messbrucke, *Elek. Mach. Tech.*, 1925, p. 263.
Ferguson, J. G., Classification of Bridge Methods, *Trans. A.I.E.E.*, Vol. 52, 861 (1934).
Wilhelm, H. T., Impedance Bridge with a Billion to One Range, *Bell Tel. Rec.*, March 1945.
Catalogs of the Leeds and Northrup Co., General Radio Co., Westinghouse Corp., and Weston Electrical Instrument Co.
General Radio Experimenter.
Definitions of the A.S.A.

WIRE LINE MEASUREMENT

By H. J. Fisher

15. TRANSMISSION MEASUREMENTS

Transmission measurements evaluate a circuit or facility (wire line, cable, radio circuit, three- or four-terminal network, etc.) in regard to its ability to transmit telegraph, voice, modulated carrier, television, or other communication signals. This evaluation usually involves two criteria, (1) effect on signal amplitude and (2) effect on signal shape. It has been found convenient to express the criteria in terms of steady-state measurements of loss or gain, phase shift, and envelope delay distortion vs. frequency, and, in addition, non-linear distortion (compression, expansion, interchannel modulation cross-talk) as a function of signal amplitude. Usually the device or circuit to be measured is one of many connected in tandem between the original signal source and the final receiver, and it is desired to know the effect on overall transmission caused by inserting the portion in question. Since the individual portions of a system are usually designed to have a nominally constant impedance (vs. frequency) of a standardized value (e.g., 600 ω, 135 ω, 75 ω) the insertion transmission can be measured directly with test equipment having the same nominal impedance. Impedance variations with frequency of the unit to be measured, of the connected circuits, or of the testing equipment, will produce errors in the insertion transmission measurement which must be considered. If all impedances are designed to have less than 5 per cent reflection coefficient these errors usually can be neglected.

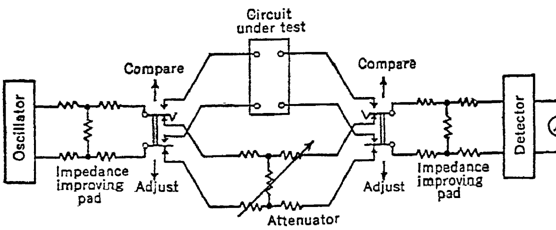


FIG. 1. Comparison Method for Measuring Insertion Loss

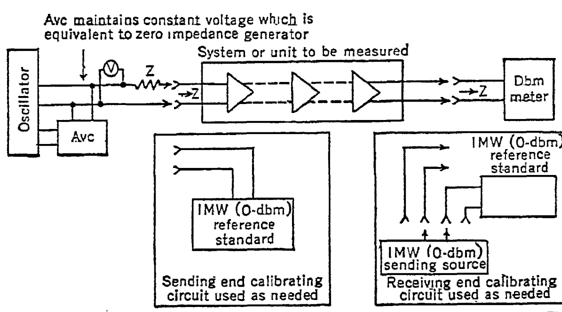


FIG. 2A. Straightaway Method for Measuring Insertion Gain or Loss

INSERTION LOSS OR GAIN. Comparison Method. Figure 1 shows a commonly used setup for measuring insertion loss. The attenuator is adjusted until equal readings are obtained on the meter for the two positions of the key. When measuring gain the attenuator is connected in tandem with the unknown. Systems or units having input frequencies that differ from the output frequencies such as modulators or mixers can be measured with this setup provided that the detector and attenuator characteristics do not change over the frequency range of interest.

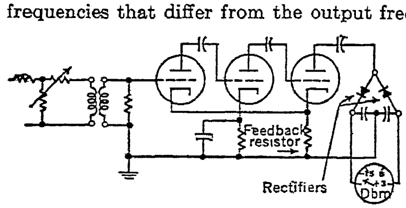


FIG. 2B. DBM Meter

Straightaway Method. Figure 2A shows a setup for making this type of measurement. A circuit of this type is required when the input and output terminals are not available at the same location. The circuit is self-explanatory except for the following items.

Dbm Meter. This is usually a broad-band amplifier having adjustable gain in 10-db steps followed by a linear rectifier (diode or varistor) with a meter calibrated in dbm over about a 13-db range. See Fig. 2B.

1 Milliwatt (0 Dbm) Reference Standard. This is usually a thermocouple circuit of the correct impedance arranged to be calibrated by standard d-c power as determined by a d-c milliammeter.

Voltage Method. For making measurements on working systems without causing disturbance to the working signals, a frequency selective voltmeter is used (see reference 3). This is a high-impedance selective heterodyne detector of adjustable sensitivity in, say, 10-db steps followed by a linear detector and dbm meter. The measurements are usually made at the pilot frequencies or other frequencies not in the signal channels. Since the working signals may be at higher levels than the pilots, it is important that the first modulator and input amplifier be operated at levels low enough to reduce the error due to intermodulation products falling at the measuring frequency. Frequency discrimination, at intermediate and final frequency stages, sufficient to eliminate all other unwanted frequencies is obtained by means of quartz crystal filters and selective interstage circuits. The required discrimination can be reduced by about 26 db by the use of a linear rectifier rather than a peak detector and about 23 db with square-law or thermal-type indicators.

MEASUREMENT OF INSERTION TRANSMISSION (LOSS OR GAIN) OF CIRCUITS AND UNITS HAVING MISMATCHED IMPEDANCES. As stated above, when the reflection coefficient of the circuit or unit being measured and the test equipment are less than 5 per cent when referred to the nominal impedance the measured insertion transmission is usually negligibly different from the true insertion transmission. In some broad-band systems the impedances of the tandem components vary considerably with frequency from the nominal, and the reflection losses incurred are sometimes employed in overall equalization of the system. When measuring components of such a system with nominal impedance test equipment there is a discrepancy between the measured and actual insertion transmission. When measuring with a voltmeter type of circuit the discrepancy is different and usually larger. This in itself causes no great complication in the maintenance of these systems since the operating limits are specified having in mind the type of test equipment which is used. Difficulty does arise when attempts are made to correlate measurements made by the two methods. A calculation using complete impedance information and rigorous insertion transmission equations is necessary.

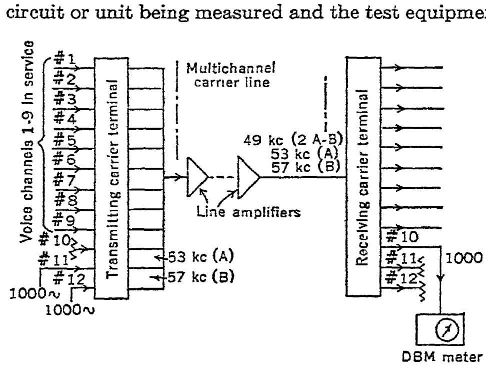


FIG. 3A. Modulation Measurements on Overall Carrier Telephone System

INTERMODULATION DISTORTION MEASUREMENT. As a result of non-linear distortion (such as occurs in vacuum-tube amplifiers) in a transmission system, frequencies other than those applied to the input of the system are produced. For example, if two tones are applied to the system, say (a) and (b), new components having frequencies such as $2a$, $2b$, $a \pm b$, $2a \pm b$, $2b \pm a$, $3a$, $3b$., etc. are produced. In some cases, these new frequencies fall outside the band of interest and need not be considered, but in wide-band systems many of these products fall within the transmitted band. In multichannel systems these new frequencies may result in interchannel interference. To prevent this interference from exceeding allowable amounts, measurement of intermodulation distortion is a necessary part of the maintenance of multichannel systems.

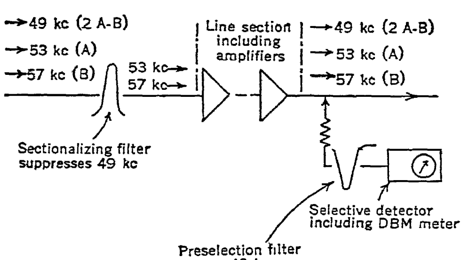


FIG. 3B. Modulation Measurements on Line Section

There are two types of test: (1) an over-all test to check whether the circuit as a whole meets specified requirements; (2) a test on a portion of the circuit or on individual repeaters to locate the defective tubes or other component or a maladjustment.

Figure 3A shows how the over-all test is made on a typical 12-channel carrier system. Channels 1 to 9 are continued in operation, channels 10, 11, and 12 being turned down. Channels 11 and 12 are energized with 1000-cycle power of specified level, and after passing through the terminal modulating equipment they appear on the line as 53 and 57 kc.

As a result of third-order intermodulation in succeeding amplifiers, 49 kc ($2a - b$) is produced. At the receiving end this product appears in voice channel 10 as 1000 cycles, where it is measured with a sensitive dbm meter. The $2a - b$ product is used for this test because it is the product most likely to be excessive. The reason for this is that the modulation component produced at each repeater tends to add in phase with the components produced by other repeaters whereas most other products add on a random basis.

To localize sources of excessive intermodulation the line may be sectionalized as shown in Fig. 3*B*. In this case, the three channels are turned down as before, the 53 and 57 kc originating as 1000 cycles at the transmitting terminal. Any 49 kc produced between the terminal and the suppression filter is suppressed and the test is essentially originated at that point. The selective detector and preselection filter are portable and can be moved as near to the suppression filter as desired, for example close enough to include only one repeater.

In other multichannel systems the modulating test tones are applied directly to working lines by means of high-frequency oscillators. They are allocated in spaces between working channels and are also chosen so that the product being measured, which may be $2a$, $3a$, $2a - b$, etc., also falls in an idle part of the spectrum.

See references 1 to 3.

INSERTION PHASE MEASUREMENT. This measurement is almost entirely restricted to the laboratory, where it is particularly useful in connection with the measurement of the feedback factor ($\mu\beta$) loop of a feedback amplifier (see references 4 and 5) and

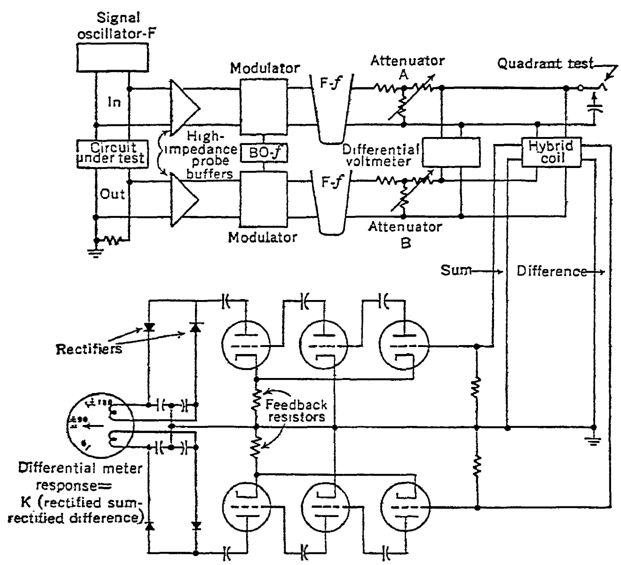


FIG. 4*A*. Direct-reading Phase-measuring Circuit

the insertion phase of networks (see reference 6). In connection with transmission of television signals over coaxial cables phase data are obtained indirectly by the integration of the envelope delay characteristic.

Figure 4*A* shows a direct-indicating type of circuit by means of which the $\mu\beta$ characteristic can be quickly obtained. Attenuators *A* and *B* are adjusted so that the input and output vectors are equal in magnitude. The loss or gain of the circuit under test can be obtained from these attenuator readings. By an alternative arrangement, using avc amplifiers this can be accomplished automatically. Assuming that vectors of equal magnitude are applied to the hybrid coil the meter reading will indicate the phase difference directly. Figure 4*B* shows the characteristics of the direct-indicating phase indicator. The difference of the rectified sum and difference outputs is used because it permits the use of a more linear portion of a cosine function $\left[\text{response} = k' \cos\left(\frac{\pi}{4} + \frac{\alpha}{2}\right)\right]$. By this means a quite linear indication of phase difference can be obtained over a 180° range and within the range of 90° ± 45° the deviation is usually negligible. This range of good linearity may be shifted to any other desired portion of the range by means of a calibrated phase shifter in one of the branches. In Fig. 4A a modulation method is shown to provide selectivity to reduce the effect of noise and to permit the use of fixed-frequency circuits, but successful broad-band sets have also been constructed. This type of set is also easily adapted to the automatic recording of $\mu\beta$.

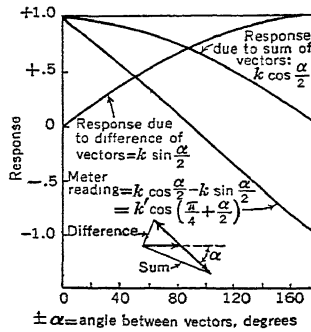


FIG. 4B. D-c Response vs. Difference in Phase of Two Equal-frequency Sine Waves

Other types of phase-measuring circuits have been devised. Generally they require the adjustment of the reference and unknown vectors to equal magnitude. Phase may then be computed from the difference in magnitude of their sum and difference, or, by means of an adjustable calibrated phase shifter (of any of several well-known types) placed in either branch, the sum or difference may be adjusted to a null and then the phase shift may be read directly from the calibrated phase shifter, due allowance being made for quadrant determination. In a variation of this method suggested by S. T. Meyers and used extensively, the sum or difference is adjusted by means of a calibrated phase shifter to equality with the reference and unknown vector. This indicates a 120° phase difference between the two applied vectors which when added to or subtracted from the indicated phase shift depending on the quadrant gives the actual phase difference.

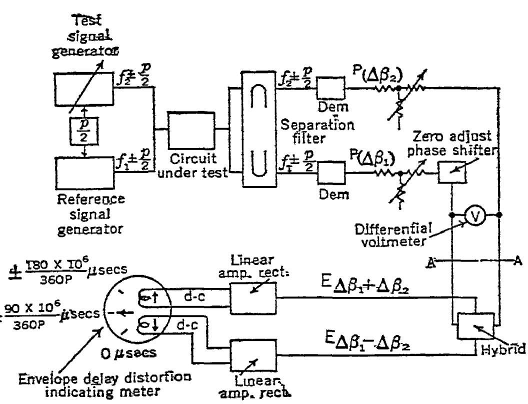


FIG. 5. Envelope Delay Distortion Measuring Circuit

Another widely used phase-measuring circuit is described in W. P. Mason's patent U. S. 1,684,403. This is known as the sum-and-difference method and requires only commonly available equipment such as an oscillator, a detector, and attenuators (see reference 6).

INSERTION ENVELOPE DELAY DISTORTION MEASUREMENT. In some types of communication, such as television, telephoto, and telegraph, distortion of signal shape is more important than for ordinary telephone communication (see references 7–12). Equally as important as attenuation distortion (vs. frequency) as a criterion for distortion of signal shape is envelope delay distortion. Envelope delay is expressed as $d\beta/d\omega$, and in an ideal system it is independent of frequency in the range of interest. Envelope delay distortion is corrected by means of phase-shift networks, the objective being to obtain a phase shift vs. frequency characteristic in the frequency range of interest having a slope $d\beta/d\omega$ which is constant. For networks and amplifiers the necessary phase data are usually obtained directly from phase measurement. For long lines it is difficult to make accurate straightaway phase measurements on account of the instability of the loss and phase of the line which may include many repeaters, and the usual practice is to measure envelope delay distortion. From these data the phase requirements for the equalizer can be computed.

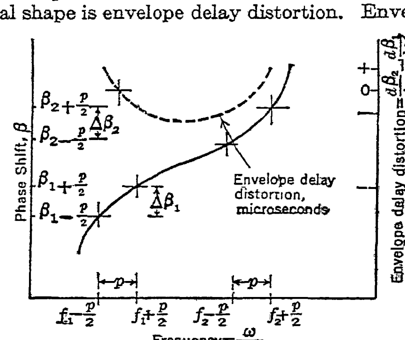


FIG. 6. Typical Phase Characteristic of a Transmission Line

Figure 5 shows a circuit for making straightaway measurements of the envelope delay distortion of long coaxial circuits. At the sending end four frequencies of equal amplitude are transmitted: $f_1 + p/2$ and $f_1 - p/2$, representing the reference signal; and $f_2 + p/2$ and $f_2 - p/2$, representing the test signal. Referring to Fig. 6, assume that the line under test has a phase vs. frequency characteristic as shown. At the distant end of the line each of the four frequencies will have been shifted in phase by different amounts $\beta_{(1-p/2)}$, $\beta_{(1+p/2)}$, $\beta_{(2-p/2)}$, and $\beta_{(2+p/2)}$. Now we can say that

$$\Delta\beta_1 = \beta_{(1+p/2)} - \beta_{(1-p/2)}, \qquad \Delta\beta_2 = \beta_{(2+p/2)} - \beta_{(2-p/2)}$$

and if p is chosen small enough then for all practical purposes

$$\frac{\Delta\beta_1}{2\pi p} = \frac{d\beta_1}{d\omega} \quad \text{and} \quad \frac{\Delta\beta_2}{2\pi p} = \frac{d\beta_2}{d\omega}$$

The envelope delay distortion $\left(\frac{d\beta_2}{d\omega} - \frac{d\beta_1}{d\omega}\right)$ is equal to $\frac{(\Delta\beta_2{}^\circ - \Delta\beta_1{}^\circ)_{10^6}}{360p}$ microseconds.

If f_2 is varied over the band of interest the measurement of $\Delta\beta_2 - \Delta\beta_1$ will represent envelope delay distortion referred to the envelope delay at the reference frequency f_1. If the set is given a "zero" adjustment by means of the "zero adjust" phase shifter so that with a distortionless line or resistance pad in place of the line under test the indicated envelope delay distortion is zero then the equipment at the receiving end will provide an indication proportional to $(\Delta\beta_2 - \Delta\beta_1)$ and therefore also to $\left(\frac{d\beta_2}{d\omega} - \frac{d\beta_1}{d\omega}\right)$, the envelope delay distortion. The receiving circuit functions as follows: the reference signal and the test signal are separated by filters as shown and then demodulated to obtain the difference products. The outputs of the two demodulators are of the same frequency p but have a phase difference equal to $\Delta\beta_2 - \Delta\beta_1$, assuming that the zero adjustment has been made. The amplitudes of the two vectors are made equal by means of the attenuators (which incidentally provides a measure of attenuation distortion) and combined in the hybrid coil from which two new vectors are obtained, one whose amplitude is a function of the sum of the two vectors $(E_{\Delta\beta_2} + E_{\Delta\beta_1})$ and the other whose amplitude is a function of the difference $(E_{\Delta\beta_2} - E_{\Delta\beta_1})$. These new vectors are rectified separately in linear amplifier rectifiers, and the difference of the d-c outputs is indicated on the two-winding zero center meter. By the use of the difference of the sum and difference d-c outputs the indication obtained is very closely proportional to $\Delta\beta_2 - \Delta\beta_1$ from 0 to $\pm180°$ and, as shown above, is therefore also proportional to $\left(\frac{d\beta_2}{d\omega} - \frac{d\beta_1}{d\omega}\right)$, the envelope delay distortion in microseconds. If the area under the plotted curve is integrated step by step by means of a planimeter or graphical methods, the phase vs. frequency characteristic can also be obtained which is the form of data required for use in designing phase-correcting networks.

By the introduction of a motor-driven signal generator and interlocking arrangements at the sending end and avc amplifiers in place of the attenuators and a recording meter at the receiving end, this circuit can be adapted to automatic recording of both the attenuation and envelope delay distortion vs. frequency characteristics.

After the initial phase correction is made by means of phase correctors in the line, the residual delay distortion is considerably reduced and it is necessary to increase the delay sensitivity of the measuring circuit to obtain data for more accurate phase equalization; this is accomplished by insertion of harmonic multipliers at A. Similarly the phase sensitivity may also be increased by increasing the frequency interval, p. The objection to this is that it does not catch the narrow interval variations.

16. NOISE MEASUREMENTS

Telephone circuit currents other than those produced by acoustic pressures on the transmitters (e.g., currents produced by electromagnetic or electrostatic induction from power circuits or from other telephone circuits, currents produced by thermal noise, vacuum-tube noise, defective components, etc.) produce noise in a telephone receiver connected to the circuit, these currents being called noise currents. Not all frequency components of noise have the same interfering effect on a telephone conversation, since the human ear and the telephone system do not respond equally to all frequencies. (See also Coordination of Communication and Power Systems, Section 10.) A measurement of the total noise power on a telephone circuit would, therefore, not be a true indication of its interfering effect.

For many years noise currents were measured by comparing the actual noise as heard in a receiver to an adjustable standard noise produced by a buzzer, the ear weighting the different frequency components of the noise. Difficulty in comparing the noise with the standard when it did not have the same frequency components, and variations between different observers when measuring the same noise, resulted in the development of meter methods, the device for measuring telephone circuit noise being known as a circuit noise meter or noise-measuring set (see references 13 and 14). This consists of a high-gain amplifier a weighting-network, a rectifier, and a d-c meter as shown in Fig. 7.

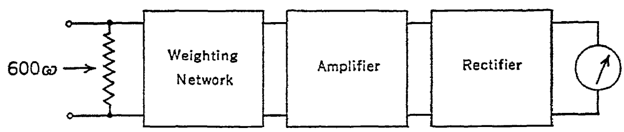


FIG. 7. Circuit Noise Meter

A typical weighting network and amplifier together have the response characteristic as shown in Fig. 8. This curve is based on a large amount of experimental data and takes into account the typical receiver-ear sensitivity and terminal trunk transmission characteristics. For lines carrying program, the weighting characteristic would be different, giving more weight to the upper and lower frequencies.

The rectifier circuit has been designed so that the d-c output for a steady-state complex input to the rectifier is proportional to the square root of the sum of the squares of the individual single-frequency voltages in the complex input. The output is a function of the average power impressed on the rectifier and not of wave shape.

A limited db scale on the meter and a gain control calibrated in decibels are provided for indicating the noise level in decibels above "reference noise," which is the term given to any circuit noise which would produce a meter reading of zero, the same reading as would be produced by sending 10^{-12} watt of 1000-cycle power into the circuit noise meter (600 ohm input).

Reference noise is defined in this manner so as to facilitate calibration and measurement. For a single type of telephone instrument and corresponding weighting network the definition is sufficient. However, there are several instruments with different frequency-response characteristics, and each requires a separate weighting network. The method of calibration makes these networks all give the same reading for a 1000-cycle input, but noise measured as equal with different networks may not be equal in interfering effect when heard with the corresponding instruments. In practice, it is customary to express noise magnitudes in dba (decibels adjusted) by adjusting the reading in db above reference noise so that equal magnitudes (in dba) represent equal interfering effects for different types of instruments. A different adjustment is required for each type of telephone instrument.

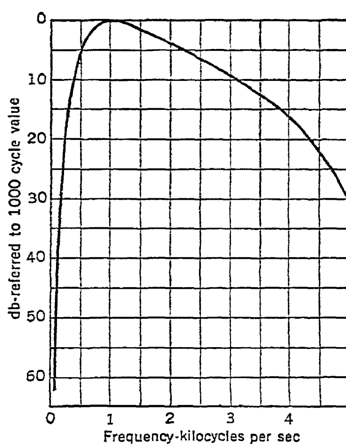


FIG. 8. Frequency Characteristic of Typical Weighting Network

The meter and associated circuits have a dynamic characteristic such that the response to sounds of short duration approximately simulates that of the ear.

17. CROSS-TALK MEASUREMENT

Transmission between separate communication circuits is called cross-talk. When the cross-talk comes principally from one other circuit, it is measured in the same way as a transmission loss by sending testing power into the disturbing circuit and measuring the cross-talk power received in the other, the ratio between the powers being expressed in decibels, since, if a circuit is not overloaded, the ratio is substantially independent of the actual power. This type of measurement is called a cross-talk coupling measurement.

In measuring the cross-talk coupling between two coterminous two-wire circuits, the generator which supplies the testing power and the receiving device which measures the received power are connected at the same ends of the circuit as in Fig. 9, the measurement being called a near-end measurement. Four-wire and carrier circuits have separate transmitting and receiving paths. For these types of circuits, it is, therefore, often desirable to connect the disturbing generator and the receiving device at opposite ends of the circuits, as shown in Fig. 10, this type of cross-talk coupling measurement being called a far-end measurement.

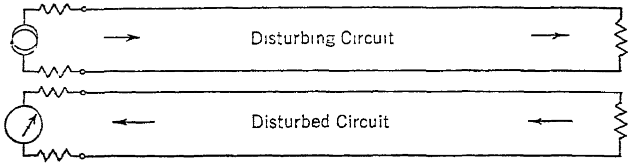


FIG. 9. Near-end Cross-talk Coupling Measurement

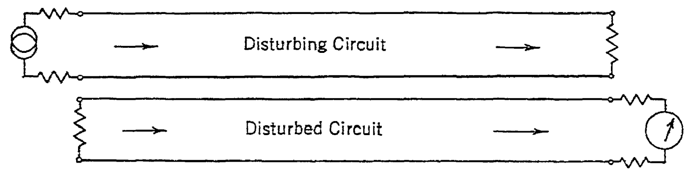


FIG. 10. Far-end Cross-talk Coupling Measurement

As the frequency of the disturbing power is changed, the cross-talk coupling between two circuits varies over a wide range. Single-frequency cross-talk measurements are therefore of little value, and generators of complex wave shape are used to obtain results approximating actual talking conditions. Either warbler oscillators or tube noise generators whose energy-frequency spectra are shaped to simulate the human voice are used as power sources. The warbler oscillator contains a frequency-changing device that causes the frequency to sweep over a wide range several times per second. The meter in the receiving or measuring device averages the results over the range.

Except when there is trouble, the cross-talk between any two circuits is generally very small. However, when there are many circuits in a group, as in telephone cables, each may produce a small amount of cross-talk in any one circuit so that the total cross-talk in that circuit may be noticeable. Coming from so many sources, it is usually unintelligible and is commonly known as "babble." The amount of babble varies with the actual volume of speech on the other circuits, being greatest at periods when the greatest number of circuits are in use.

Since the sources of babble are numerous, and since the volume on the disturbing circuits cannot be controlled, the babble noise on a circuit must be measured in some such manner as speech volume, or noise, the measurement being one of power, rather than a power ratio measurement as in cross-talk coupling.

Tests are made by connecting a high-impedance vacuum-tube measuring device across an idle circuit. This receiving device is similar to that used in transmission level or volume tests with the exception that the amplifier does not have a uniform frequency response, the characteristic of the amplifier being similar to that used in noise measurements. If observations with this device for short periods during several successive days show no abnormal conditions, the circuit is considered satisfactory. A high-impedance device is employed so that it will not interfere with normal use of the circuit.

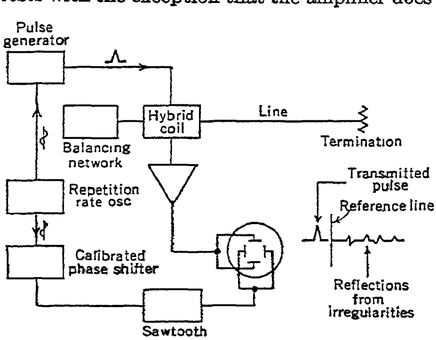


FIG. 11. Echo Indicator

18. ECHO TESTING OF LINES

This method of test is an adaptation of radar methods and has been used in two principal applications: (1) location of gross faults in cable and open-wire circuits, and (2) detection of irregularities of coaxial cable circuits used for television.

Figure 11 shows a circuit applicable to this test. The fundamental circuit is similar for both applications, the differences being in the shape and length of the transmitted pulses, the repetition rates, and the sweep speeds (see references 12, 15, and 16).

resistance and reactance curves are no longer smooth but change periodically, curve B showing the effect of an irregularity on the effective resistance. The reactance changes in a similar manner. This effect is utilized in locating the irregularity, as there is a relation between the separation of peaks on the curve and the distance to the irregularity.

When an alternating current strikes a circuit irregularity such as a sudden change in impedance, some of the current is reflected towards the sending end, the amount so reflected depending upon the size of the irregularity. Sometimes this reflected current aids the current entering the line and sometimes it reduces it, depending upon the distance to the irregularity and the frequency. When the distance from the sending end of the line to the irregularity is great a larger number of wavelengths is included between the two points than when the line is short; consequently a smaller change in frequency is necessary to add an extra wavelength. Each wavelength, half of which adds to and half of which subtracts from the original current, thereby decreasing or increasing, respectively, the impedance of the circuit, causes a peak or hump in the curve.

As an example of the manner in which the distance to an irregularity can be determined (see Fig. 12), let d be the distance to an irregularity. The reflected current must travel from the sending end to the irregularity and back again to the starting point, so that it really travels twice the distance or $2d$. If the length of one wave is W_1, the total number of wavelengths in the reflected current equals twice the distance divided by the wavelength. Let the number of waves be designated by N; then $N = 2d/W_1$ at some particular frequency which can be called f_1. Assume that f_1 is the frequency corresponding to one of the peaks on the impedance curve.

As brought out in the foregoing, when the frequency has been increased so that one more wavelength is included in the double length of line, another peak will be produced in the impedance curve. Let this frequency be designated as f_2. There are now $N + 1$ wavelengths at a frequency f_2, or $N + 1 = 2d/W_2$. The distance d to the irregularity has not changed, but the length of one wave has changed to some value W_2.

As the wave travels along a telephone line at practically the same velocity at all frequencies, the wavelength can be expressed in terms of velocity and frequency. If an alternating current flows over a circuit at some velocity V miles per second, the length in miles of any one wave, W, is equal to the velocity divided by the number of waves per second, or the frequency. In other words, $W = V/f$.

As shown above, the number of waves in the double path of the reflected current is equal to N for frequency f_1 and $N + 1$ for frequency f_2. In turn $N = 2d/W_1$ and $N + 1 = 2d/W_2$. Since the wavelength equals the velocity in miles divided by the frequency, the wavelength for any particular frequency such as f equals the velocity divided by that frequency. Therefore,

$$W_1 = \frac{V}{f_1}$$

and

$$W_2 = \frac{V}{f_2}$$

Substituting these values of W_1 and W_2 in the equations

$$N = \frac{2d}{W_1}$$

and

$$N + 1 = \frac{2d}{W_2}$$

respectively. Then

$$N = \frac{2d/V}{f_1} = \frac{2df_1}{V}$$

and

$$N + 1 = \frac{2d/V}{f_2} = \frac{2df_2}{V}$$

The frequency f_1 represents one peak on the curve, and f_2 represents the next peak as the frequency increases. From the curve can be determined the number of cycles difference in the two frequencies representing adjacent peaks, this difference, of course, being equal to $f_2 - f_1$. Combining the two equations above so that the term $f_2 - f_1$ will be present, subtract N from $N + 1$. Then

$$N + 1 - N = \frac{2df_2}{V} - \frac{2df_1}{V}$$

This gives

$$1 = \frac{2df_2}{V} - \frac{2df_1}{V}$$

or

$$V = 2d(f_2 - f_1)$$

From this the distance to any irregularity may be determined provided the difference in frequency between two adjacent humps and the velocity with which an alternating current flows along the line are known. The above equation may be written as

$$d = \frac{V}{2(f_2 - f_1)}$$

Expressed in words this means that the distance to any irregularity equals the velocity with which an alternating current flows along the line in question divided by twice the difference in frequency between adjacent humps.

The velocity of transmission for all types of circuits is determined experimentally by introducing a known irregularity and solving the equation for V. In practice it is usual to take the difference in frequency between several adjacent peaks, between approximately 700 and 1500 cycles, and use the average difference in velocity at different frequencies which causes the interval between peaks to change slightly with frequency.

Measurements are usually made with the simple form of impedance bridge shown schematically in Fig. 12.

21. D-C AND LOW-FREQUENCY LINE TESTING

Telegraph and telephone line test boards are generally equipped with special types of voltmeters and Wheatstone bridges for making periodic tests of the line wires and for the location of faults. These faults are of three general types: grounds, crosses, and opens; and they may have any value of resistance from zero to several megohms. The voltmeter provides a simple method of determining the type and magnitude of a fault, as shown in Figs. 13, 14, 15, and 16.

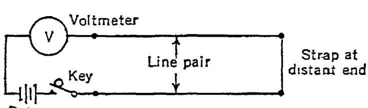


FIG. 13. Continuity Test

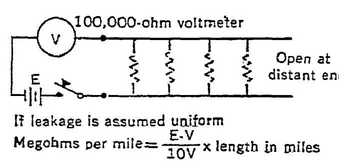


FIG. 14. Insulation Test

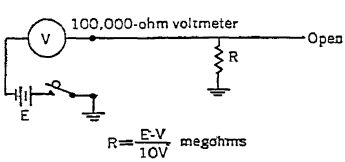
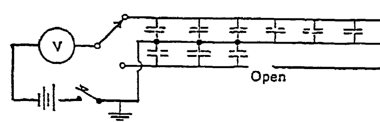


FIG. 15. High-resistance Leak to Ground



When key is closed the charging current will cause a momentary deflection. The duration is proportional to the capacitance. Thus by comparing the deflection obtained with the defective wire with that obtained with a good conductor of known length the distance to the fault can be estimated.

FIG. 16. Voltmeter Test for Open

For the actual location of faults the Wheatstone bridge is used as illustrated in Figs. 17, 18, 19, 20, and 21. In these figures L is the length of the line and d is the distance

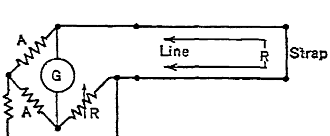


FIG. 17. Simple Bridge to Measure Loop Resistance. At Balance $R_L = R$.

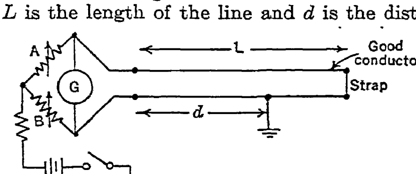
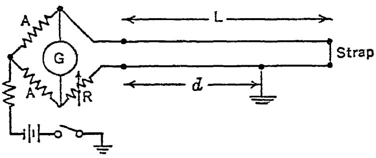


FIG. 18. Simple Murray Loop Test for Grounded Conductor. When A and B are adjusted for balance

$$d = 2L\frac{B}{A+B}$$

from the testing end to the fault. It is assumed in the illustrations that the good and faulty conductors used in the bridge measurement have the same resistances per unit length in the case of grounds and crosses, and the same capacitances per unit length in the case of opens.

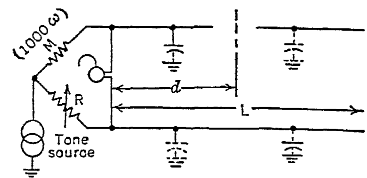


Note: The conductor unit resistance will not usually be accurately known. For this reason a fault is ordinarily located on a percentage basis by making both simple loop and Varley measurements.

FIG. 19. Simple Varley Loop Test for Grounded Conductor

$$d = L - \frac{R}{2K}$$

(K = ohms per conductor unit length.)



When R is adjusted to give minimum response in telephone receiver

$$d = \frac{R}{M} L$$

FIG. 20. Simple Murray Loop Test for Opens. [Suitable only for short lines (1/2 mi cable). For longer lines use Fig. 21.

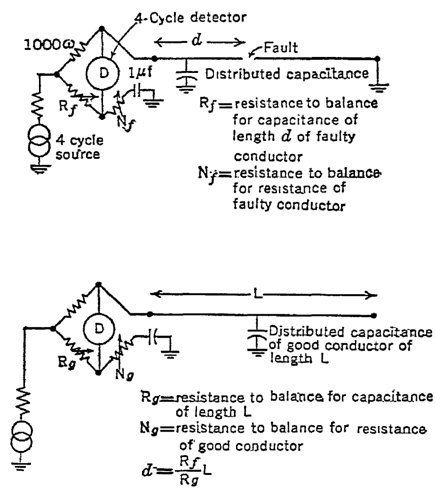


FIG. 21. To Locate an Open, a Bridge Reading Is Taken on the Open Conductor and Compared with the Reading Obtained on a Good Conductor of Known Length

If the fault is a cross between wires, the second crossed wire is substituted for the ground connection to the battery key.

It is important that the location be as accurate as possible to lessen the overall fault clearing time, and to this end variations of the simple bridge tests are used to minimize errors (see references 22–29).

BIBLIOGRAPHY

1. Bennett, W. R., Cross-modulation Requirements on Multichannel Amplifiers below Overload, *Bell Sys. Tech. J.*, Vol. 19, 587–605 (October 1940).
2. Kinzer, J. P., Measurement of Modulation in Carrier Amplifiers, *Bell. Labs. Rec.*, August 1941.
3. Tidd, Rosen, and Wenk, New Test Equipment and Testing Methods for Cable Carrier Systems, *Trans. A.I.E.E.*, Vol. 66 (1947).
4. Bode, H. W., *Network Analysis and Feedback Amplifier Design*, D. Van Nostrand, New York.
5. Black, H. S., Stabilized Feedback Amplifier, *Bell Sys. Tech. J.*, Vol. 13, 1–18 (January 1934).
6. Alsberg and Leed, A Precise Direct Reading Phase and Transmission Measuring System for Video Frequencies, *Bell Sys. Tech. J.*, Vol. 28, 231–238 (April 1949).
7. Elliot, J. S., Precise Measurement of Insertion Phase Shift, *Bell Labs. Rec.*, Vol. XVI, No. 8, 285.
8. Mead, S. P., Phase Distortion and Phase Distortion Correction, *Bell Sys. Tech. J.*, Vol. 7, 195–224 (April 1928).

9. Lane, C. E., Phase Distortion in Telephone Apparatus, *Bell Sys. Tech. J.*, Vol. 9, 493–521 (July 1930).
10. Nyquist, H., and Brand, S., Measurement of Phase Distortion, *Bell Sys. Tech. J.*, Vol. 9, 522–549 (July 1930).
11. Wentz, J. F., Measuring Transmission Speed of the Coaxial Cable, *Bell Labs. Rec.*, June 1939.
12. Strieby, M. E., and Weis, C. L., Television Transmission, *Proc. I.R.E.*, Vol. 29, 300–321 (July 1941); also p. 381.
13. Barstow, J. M., Blye, P. W., and Kent, H. E., Measurement of Telephone Noise and Power Wave Shape, *Elec. Eng.*, Vol. 54, 1307–1315 (December 1935).
14. Castner, T. G., Dietze, E., Stanton, G. T., and Tucker, R. S., Indicating Meter for Measurement and Analysis of Noise, *Trans. A.I.E.E.*, September 1931, pp. 1041–1047.
15. Schott, J. T., The Lookator, *Bell Labs. Rec.*, October 1945, p. 379.
16. Abraham, Lebert, Maggio, and Schott, Pulse Echo Measurements on Telephone and Television Facilities, *Trans. A.I.E.E.*, Vol. 66, pp. 541 to 548.
17. Swift, G., Amplifier Testing by Means of Square Waves, *Communications*, February 1939.
18. Transient Response of a Broadcast System, *General Radio Experimenter*, April 1940.
19. Network Testing with Square Waves, *General Radio Experimenter*, December 1939.
20. Williams, J., Square Wave Testing of Amplifiers, *Radio News*, January 1944, p. 24.
21. Ferris, L. P., and McCurdy, R. G., Telephone Circuit Unbalances, *J. A.I.E.E.*, Vol. XLIII, No. 12 (December 1924).
22. Palmer, W. T., A-c Method of Fault-localization in Telephone Cables, *P.O.E.E. J.*, Vol. 23, No. 1 (April 1930).
23. *Trans. A.I.E.E.*, Vol. 43, 423 and 1320.
24. Northrop, E. F., *Methods of Measuring Electrical Resistance*, McGraw-Hill, 1912.
25. Ritter, E. S., Cable Testing, *Printed Paper* 104, Institute of P.O.E.E., 1923.
26. Edwards, P. G., and Herrington, H. W., The Location of Opens in Toll Telephone Cables, *Bell Sys. Tech. J.*, Vol. 6, No. 1 (January 1927).
27. Henneberger, T. C., and Edwards, P. G., Bridge Methods of Locating Resistance Faults in Cable Wires, *Bell Sys. Tech. J.*, Vol. 10, 382–407 (July 1931).
28. Metson, G. H., An Accurate Method of Sub-localizing Cable Faults, *P.O.E.E. J.*, Vol. 30, No. 2, 99 (1937).
29. Allan, J. M., Two Methods of Locating Cable Faults, *P.O.E.E. J.*, Vol. 29, Part 2 (July 1946).

ROUTINE MEASUREMENTS ON A-M AND F-M BROADCAST RECEIVERS

By W. O. Swinyard

In receiver measurements it is sometimes convenient to use the logarithm of the stage gain so that gains can be added, instead of multiplied, to determine the overall non-regenerative gain, and to express these in terms of decibels. It is also convenient to discuss overall sensitivities in terms of decibels below 1 volt, in which case the microvolt sensitivity is used to determine a voltage ratio. Though this is strictly a misuse of the term, it is convenient and will lead to no serious difficulty as long as the user is aware of the limitations of these practices.

22. OVERALL A-M RECEIVER MEASUREMENTS

REQUIRED TEST EQUIPMENT. Standard-signal generator, standard dummy antenna, standard test loop, output wattmeter, audio-frequency generator, distortion meter or wave analyzer, and an auxiliary 1000-kc signal generator.

STANDARD TEST CONDITIONS. Standard line voltage is 117 volts for a-c, d-c, and a-c/d-c receivers. Receivers designed for a-c and d-c operation are usually tested on alternating current and check measurements are made on direct current.

Measurements on automobile receivers should be made using a battery which provides 6.6 volts at the receiver battery terminals.

The normal test voltage for receivers designed for farm lighting systems is 36 volts.

Battery-operated receivers should be tested using new batteries of the type and voltage specified by the receiver manufacturer.

The volume control and the tone control or controls should be set to provide maximum 400-cycle output. If a selectivity control is provided, it should be set, for the initial tests, to provide greatest selectivity. The effect of the tone and selectivity controls on the performance should be determined by special tests.

RECEIVER ALIGNMENT CONDITIONS. The overall sensitivity, selectivity, and range coverage of the receiver are first measured without disturbing the receiver alignment. These measurements are followed by the single-stage measurements, after which the receiver is aligned, in accordance with the manufacturer's service instructions if available, and complete overall measurements are made. Normally, no overall measurements are made with the receiver aligned at each test frequency. However, in certain cases such measurements might be desirable since they would show the effect of circuit misalignment on the overall sensitivity, assuming no regeneration. In all cases, however, the frequencies at which the circuits are in exact alignment should be noted.

Figure 1 shows the equipment as it is set up for overall measurements.

RANGE COVERAGE. The maximum and minimum frequencies to which the receiver can be tuned in each band are recorded for the "as received" and later for the aligned conditions.

OVERALL SENSITIVITY (SENSITIVITY-TEST INPUT). This test normally consists of determining the sensitivity-test input at three to six points in each wave band. The output of the standard-signal generator is fed into the input terminals of the receiver through the standard dummy antenna. The output meter is connected to the secondary of the output transformer, and the load is adjusted to the proper value. The receiver is then turned on and the voltage of the power source set to the correct value. After it has throughly warmed up it is tuned to the modulated signal-generator output and the controls are adjusted to provide maximum 400-cycle output. The 400-cycle filter is then switched in to remove the noise, and the sensitivity-test input is determined. It may be measured in decibels below 1 volt or in microvolts.

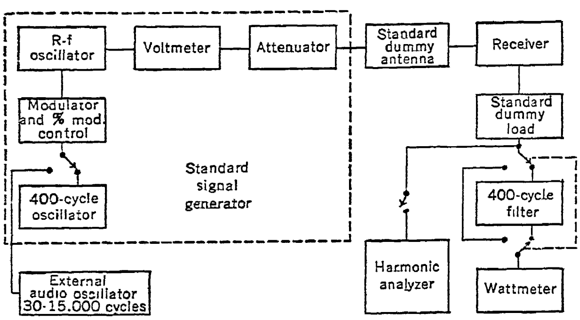


FIG. 1. Arrangement of Apparatus Used for Overall Measurements

If the receiver employs a selectivity control its effect on the sensitivity is determined by additional measurements which are usually made at 600 kc, 1000 kc, and 1400 kc.

The sensitivity of battery-operated receivers is measured in the normal way and with A and B batteries whose terminal voltage has dropped to 1.1 volts for the A and to 60 per cent of the nominal value for the B battery.

A 400-cycle filter used with the output wattmeter usually will reduce the noise voltage in the output to a negligible level. However, there may be cases where noisy or extremely sensitive receivers are being measured when allowance must be made for residual noise.

EQUIVALENT-NOISE-SIDE-BAND INPUT. The equivalent-noise-side-band input is taken equal to the input of a single side band of 400-cycle modulation which will produce an output from the receiver equal to the noise output, other conditions being the same. The reason noise of both side bands has to be identified with a single-side-band component of modulation is that there is a random-phase relation between the noise side bands as distinguished from the specific phase relations existing between the carrier and each pair of side-band components of modulation. The equation for *ENSI* is:

$$E_n = 0.3E_s \frac{E_n'}{E_s'}$$

where E_n is the desired *ENSI* in microvolts, E_s is the carrier level in microvolts, E_n' is the rms output voltage of the noise alone with an unmodulated carrier applied to the antenna, and E_s' is the output voltage due to modulating this carrier 30 per cent at 400 cycles. The coefficient 0.3 is required because the test input is modulated 30 per cent, and only one side band is considered.

To obtain *ENSI*, the sensitivity is first determined making due allowance for noise. This gives a convenient value for E_s in microvolts. The corresponding 400-cycle output voltage gives the value of E_s'. The modulation is then removed, the 400-cycle filter switched out, and the rms voltage due to the noise is noted. This gives the value of E_n' if a thermocouple meter is used or if the proper correction is applied for the particular meter used. The corresponding *ENSI* may then be computed from the equation given above.

The value of E_s' and E_n' may be observed at a higher signal input level since only their ratio enters. A good practice is to increase the carrier input at which *ENSI* measurements are made, whenever the ratio E_n'/E_s' exceeds 1. This makes the effective increase in noise power due to beats between noise side bands negligible relative to the power due to beats between the noise side bands and the carrier.

It should be noted that it is possible to determine *ENSI* without using a 400-cycle filter. Where this procedure is followed, it is convenient to set the attenuator to the point which makes the noise-output power equal to the signal-output power. *ENSI* is then the signal input multiplied by the per cent modulation.

SELECTANCE RATIOS. It is of interest to know how well the receiver discriminates against signals located 10 kc away on either side of the desired channel. The regular overall sensitivity setup is used. The selectance ratio in decibels is the difference between the signal input in decibels below 1 volt required for normal test output on the channel to which the receiver is tuned and that required for normal test output on the adjacent channel. The selectance ratios are usually measured at 600 kc, 800 kc, 1000 kc, and 1400 kc in the broadcast band and at corresponding points in the long-wave band if one is used.

IMAGE RATIOS. The setup for these measurements is the same as that for overall sensitivity. After measuring the sensitivity, the receiver tuning is left undisturbed and the signal generator is set to the image frequency. The least signal-input voltage with the signal modulated 30 per cent at 400 cycles required for normal test output is the image sensitivity. The image ratio in decibels is the difference between the overall and image sensitivities, when both are expressed in decibels below 1 volt.

There are other spurious responses such as the half i-f image which are due to oscillator harmonics beating against either the fundamental of an interfering signal or harmonics of that signal which may be generated in the receiver. However, these are not usually measured in a routine receiver analysis.

I-F REJECTION RATIOS. The sensitivity of the receiver to the intermediate frequency is determined with the regular setup used for overall-sensitivity measurements (including the standard "all-wave" dummy antenna). The i-f rejection ratio in decibels is the difference between overall and i-f sensitivities in decibels below 1 volt. Measurements are usually made at three points in each wave band. In the broadcast band these points are usually the two lowest and the highest frequency test points.

OVERALL SELECTIVITY. The overall selectivity is measured at the center of the band (1000 kc in broadcast receivers). The setup is the same as that used for overall-sensitivity measurements, and the measurement can conveniently be made after the measurement of sensitivity and *ENSI* at 1000 kc. The total widths of the selectivity curve at points 6, 20, 40, 60, and 80 db down from the resonant voltage peak are recorded in kilocycles as W_6, W_{20}, W_{40}, W_{60}, and W_{80}. The procedure is as follows: the receiver is tuned to resonance at 1000 kc and the attenuator set to the sensitivity-test input; the signal generator is then tuned off resonance and the output voltage is increased 6 db; the generator is tuned toward resonance until normal test output is secured and the frequency is noted; the generator is then tuned through resonance to the point on the other side where normal test output is secured and the frequency is again noted. The difference in kilocycles between these two readings is W_6. For W_{20}, W_{40}, etc., the process is repeated. At this point it should be pointed out that the modulation should be removed while the generator is tuned through resonance to avoid damage to the output meter and that care should be taken to avoid errors in the measurement due to possible back-lash in the signal generator frequency-dial mechanism.

Where variable i-f selectivity is employed, selectivity measurements are usually made at each position of the selectivity control switch, or at three positions, in the case of continuously variable selectivity, at the extremes and middle settings of the control.

An additional measurement of selectivity is made in the case of battery-operated receivers, namely, under the "dead battery" conditions previously described.

For a receiver employing a long-wave band, it is desirable to measure the selectivity at the midpoint of this band since the normal selectivity of low-frequency preselector circuits may result in appreciable narrowing of the skirt selectivity.

In the case of a receiver incorporating a double avc system normal selectivity measurements will indicate sharper selectivity than is actually provided by the selectivity of the resonant circuits. This is true to some extent in any receiver which has less selectivity to the avc detector than to the signal-frequency detector. In some cases it may be advisable to disable the avc before making selectivity measurements.

WHISTLE MODULATION. When a signal having a frequency close to twice the intermediate frequency is being received it is often accompanied by an audible whistle or tweet. This is measured at several values of signal-input level and compared with the corresponding 400-cycle modulated signal by expressing the whistle in terms of the per cent modulation at 400 cycles required to give an output voltage equal to that resulting from the whistle.

Measurement of the equivalent whistle modulation is made with the regular setup for overall-sensitivity measurements. The signal generator with the output set to the appropriate value and with zero modulation is tuned to approximately twice the intermediate frequency. The exact setting is chosen by slowly turning the generator from about 30 kc below to 30 kc above the frequency that is twice the intermediate frequency while the receiver tuning control is rocked. The generator is set at the point providing the greatest whistle output. The receiver, not the generator, is detuned to produce a whistle of maximum intensity to the ear, the volume control being set so that the output voltage at this point is well below the overload level. The 400-cycle filter is then switched in; the modulation is applied to the signal generator and adjusted until the output approaches as closely as possible the output previously noted. The ratio of the 400-cycle output voltage to the whistle voltage multiplied by the per cent modulation gives the whistle modulation in per cent. This procedure is followed for inputs of 0, 20, 40, and 60 db below 1 volt.

The above procedure must be modified when measuring the whistle modulation at input levels of 80 db below 1 volt and less, since the noise tends to mask the whistle when the filter is not used. Therefore, the 400-cycle filter is switched in and the whistle carefully adjusted to give maximum output. Then the receiver is tuned slightly to one side of the carrier, the modulation applied, and the whistle modulation determined as before. The equivalent whistle modulation measured at 400 cycles is usually substantially less than that of higher frequency whistles, when appreciable avc voltage is developed. However, these low-level measurements are usually only important qualitatively, and they indicate the need for work on the receiver to remove their causes. In thoroughly shielded receivers of correct design the whistle is usually not measurable with inputs below 80 db below 1 volt.

OUTPUT AND AVC CHARACTERISTICS. These measurements are made in the middle of the band (1000 kc for broadcast receivers) using the setup and test conditions for overall sensitivity. The signal is modulated successively at 0, 10, and 30 per cent for each value of signal input. The signal input is varied from 120 db to 0 db below 1 volt, and the audio output for each of the three modulation percentages is plotted against the corresponding signal input. The 400-cycle filter is switched in to remove the noise from that portion of the 30 per cent curve which is below the overload level, and the data are recorded with and without the filter for this portion of the curve. The 0 per cent output curve indicates the noise output of the receiver at full sensitivity. It also indicates the presence of hum modulation and motorboating at high signal-input levels.

The avc characteristic is taken in the same manner as the 30 per cent modulation curve except that the volume is reduced sufficiently to prevent overloading the output amplifier. Usually a reduction in the output voltage of 6 db with a 0 db below 1 volt signal input is sufficient. The 400-cycle filter should be used where a measurable amount of noise is present.

AVC FIGURE OF MERIT. The avc figure of merit can be obtained from the avc characteristic. It is the number of decibels decrease in signal input necessary to reduce by 10 db the output obtained at a signal input level of 20 db below 1 volt.

TWO-SIGNAL SELECTIVITY (CROSS-TALK INTERFERENCE). For these measurements the generators may be coupled to the receiver under test in either a series or parallel arrangement. The measurement is made at 1000 kc for two values of desired signal input: 46 db and 0 db below 1 volt for home receivers and 46 db and 14 db below 1 volt for automobile receivers. The 1000-kc signal is tuned in with the volume control set to provide a medium-strength a-f output from a 1 per cent modulated signal. Modulation is then removed; the two generators are connected together with the interfering signal modulated 30 per cent and introduced at amplitudes which result in the output previously noted for the 1 per cent modulated desired signal. The interfering signal measurements are taken at every channel on both sides of 1000 kc up to a 100-kc difference unless the necessary signal input reaches the maximum output of the generator before the plus and minus 100-kc points have been reached.

If an auxiliary 1000-kc generator is used, the necessary output adjustments can be made using the standard-signal generator modulation for both desired and interfering signals. When the two are connected in series for the test the auxiliary 1000-kc generator, having no modulation, serves as the desired signal.

It should be noted that the impedances of the two standard dummy antennas which are necessary if the signal generators are connected in parallel should be double the normal values. If the two generators have equal output impedances independent of the attenuator settings, the effective output voltage produced by either one is only half the indicated output.

In making the two-signal selectivity tests two whistles will usually be encountered on the low-frequency side of 1000 kc. If the intermediate frequency is 455 kc, the first occurs at approximately 955 kc and is due to the beat between the fundamental of the receiver oscillator, 1455 kc, and the second harmonic of the input signal. The second occurs at approximately 910 kc and is the regular twice-i-f whistle to which previous reference has been made.

HARMONIC DISTORTION. A distortion analysis is made at 1000 kc using the setup for overall sensitivity in conjunction with a distortion meter or a wave analyzer. Three sets of measurements of per cent harmonics are made. For the first the signal input is maintained constant at 46 db below 1 volt, with 30 per cent modulation at 400 cycles. Harmonics are measured at several output levels up to and including the maximum obtainable. It is desirable to choose one output level so that 10 per cent total distortion results since this is usually chosen as representing the maximum undistorted output. For the second set of measurements the signal input is maintained constant at 46 db below 1 volt and the receiver volume control is adjusted for normal test output. The modulation is then set to 10 per cent, 50 per cent, and 80 or 100 per cent, and the harmonics are measured for each value of modulation percentage. The final measurements are made keeping the modulation at 30 per cent, the output at normal test output, setting the signal-input level to 60 db, 40 db, 20 db, and 0 db below 1 volt and measuring the harmonics at each value of signal input. These values of signal input may have to be modified for receivers of special types. If a wave analyzer is used the total per cent harmonic distortion is calculated as the square root of the sum of the squares of the individual per cent harmonics.

OVERALL ELECTRICAL FIDELITY. The regular overall-sensitivity-measurement setup is used for this measurement except that the signal generator is modulated 30 per cent by an a-f signal generator of variable frequency and the 400-cycle filter is switched out. The receiver is first tuned to 1000 kc using a weak signal. The signal input is then set to 46 db below 1 volt, and the receiver volume control is set to provide an output well below overload. The modulation frequency is then increased until the output is reduced by about 14 db, and the tuning control is finally adjusted for minimum output at this frequency. The modulation is then set back to 400 cycles and a final adjustment of the output voltage is made. The output is then measured at a sufficient number of modulation frequencies to permit plotting a curve showing the relative response, in decibels, versus modulation frequency using the 400-cycle output voltage as a reference. Curves are made showing the maximum overall fidelity and the effect of the tone control on the fidelity. Usually two curves are recorded: (1) with the tone control set for maximum highs, and (2) with the tone control set for minimum highs. In the case of 1.4-volt battery-operated receivers the fidelity is also measured using "dead batteries." If a bass-compensated volume control is used, measurements should be made with the arm set both above and below the tap to show the effect of the bass compensation. It may be better to disconnect the network from the tap in order to avoid overload.

TESTS ON PUSH-BUTTON TUNERS. Receivers provided with a mechanical push-button tuning mechanism are subjected to tests devised to show how accurately the push buttons can be set and how well they return to the frequency to which they are set. For these tests the buttons are all set up to tune a signal near the high-frequency end of the band to zero beat with a signal from another generator which is set to the intermediate frequency. For each push button the gang is opened, the button actuated, and the resultant beat recorded as a frequency error. This is repeated several times, usually eight, and similar tests are made from the closed position of the gang. Finally tests are made with the gang alternately opened and closed before the button is pushed. All measurements for each button (24) are then averaged taking account of the sign, and the result is recorded as a setting error. The setting errors for each button are averaged without regard to sign to secure the average setting error. The deviations of all the errors for each button from the figure representing the setting error for that button averaged without regard to sign give the mean tuning deviation. The average mean tuning deviation is the average of the mean tuning deviations.

MEASUREMENTS USING STANDARD TEST LOOP. The following measurements which employ the standard test loop (Fig. 2) are made on loop receivers: sensitivity,

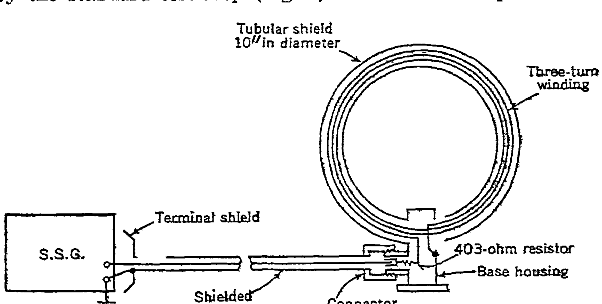


FIG. 2. Test Loop

ENSI image, i-f and selectance ratios, and loop figure of merit. The test loop is connected to the output of the signal generator and set up 24 in. away from the receiver loop with the loops arranged coaxially. The overall sensitivity in decibels below 1 volt per meter is 20 db below the signal generator attenuator reading when the input is adjusted for normal test output. The loop figure of merit is the difference between the sensitivity in decibels below 1 volt per meter and the sensitivity at the first grid measured in decibels. It is recorded as minus when the absolute magnitude of the overall sensitivity in decibels below 1 volt per meter is less than the first grid sensitivity in decibels. Where no external antenna connection is provided other measurements than these must of necessity be made using the loop input. In all cases the same procedure is followed as has been previously outlined except that the input is specified in decibels below 1 volt per meter. Many signal generators operated with the standard test loop will not provide an input to the receiver loop of more than 20 db below 1 volt per meter when the test loop is operated under normal conditions.

AUDIO FEEDBACK FACTOR. If the audio amplifier of the receiver employs negative feedback a measurement of overall a-f gain is made with the feedback removed, and the increase in gain measured in decibels at 400 cycles is expressed as the audio feedback factor.

OSCILLATOR DRIFT. This test may be carried out as follows: The receiver is tuned to a secondary frequency standard or to a signal generator of known stability characteristics. The resultant beat produced as the oscillator drifts is measured by zero beating it against an audio signal from an amplifier fed by an a-f generator. Oscillator drift measurements are made at some point near the high-frequency end of each band. The receiver is tuned to about the right point and then turned on and as quickly as possible tuned to zero beat against the frequency standard or signal generator. Measurements are made every 5 minutes for a half hour and then every 15 minutes for 2 hours or until the oscillator frequency has become stabilized. If the oscillator drift is such that the beat note goes above the range of audibility the drift can be determined directly by zero beating with the signal generator for each measurement. The direction of drift can be determined by noting the effect on the beat note of a small change in the capacitance in the oscillator tank circuit—a change such as is produced by bringing a finger up close to the tuning condenser stator.

23. SINGLE-STAGE MEASUREMENTS

REQUIRED TEST EQUIPMENT. Standard-signal generator, standard dummy antenna, output wattmeter, vacuum-tube voltmeter, tuning wand, reactance meter or *Q* meter, high-resistance voltmeter, and a wattmeter.

HIGH-FREQUENCY MEASUREMENT PRECAUTIONS. In many cases measurements cannot be made properly by substituting a vacuum-tube voltmeter for a tube or by placing it in parallel with a tube. The inherent capacitance of the stage may be so low that a change in its value may decidedly influence the shape of the selectivity curve. In such cases it is suggested that the procedure to be described later for use in making single-stage measurements on f-m receivers be followed.

ANTENNA GAIN AND BAND WIDTH. The voltage gain from the antenna to the first grid is measured at all the test frequencies used for sensitivity measurements. The width of the resonance curve 6 db down is measured at three points in each wave band. For these measurements the signal generator output voltage unmodulated is fed into the receiver antenna circuit through the standard dummy antenna. The voltage developed across the output of the antenna circuit is measured by means of a vacuum-tube voltmeter to which the lead normally connected to the grid of the first tube is connected. The signal generator output voltage is increased by 6 db and the generator is then detuned on each side of resonance until the output drops to the reference value. If there are two tunable circuits ahead of the first tube the gain and band widths are also measured on the first circuit alone. Where measurements are made at high frequencies, the precautions previously noted must be observed.

COIL MISALIGNMENT. When there are two tunable circuits ahead of the first tube or an antenna and an r-f stage, the misalignment of these circuits can be measured by means of a reactance meter or a *Q* meter. For a receiver with a tuned r-f stage, the receiver tuning condenser is set at a minimum capacitance and the standard dummy antenna is connected across the antenna-ground terminals of the receiver. The r-f circuit is used as a standard and connected across the condenser terminals of the *Q* meter, with a standard coil covering the desired range plugged into the coil terminals. The frequency of the *Q* meter is set to the high-frequency alignment point and the *Q* meter condenser adjusted for maximum *Q* indication. The antenna circuit is then substituted for the r-f circuit and the antenna trimmer adjusted to bring the antenna circuit into exact alignment with the r-f circuit as indicated by a maximum reading on the *Q* indicator.

After the antenna and r-f circuits have been aligned, the Q meter is set to each of the test frequencies and at each point the r-f circuit is connected to the condenser terminals and the Q meter tuning condenser is adjusted for maximum Q indication. The antenna circuit is then substituted for the r-f circuit and the Q meter condenser is readjusted for maximum Q indication. The difference between the two condenser settings is the misalignment correction in micromicrofarads between the two circuits.

Another measurement is made using the Q meter to determine the effect changing the antenna capacitance has on the broadcast or long-wave band tuned antenna circuit. For these measurements the receiver tuning condenser is set to minimum capacitance and the condenser terminals of the Q meter are shunted across the tuned antenna circuit. A dummy antenna, identical to the standard dummy antenna except that the series capacitance element is adjustable, is set to the standard value of 200 $\mu\mu$f and connected across the antenna input circuit. The Q meter is set to 500 kc, and the Q meter condenser is adjusted for maximum Q indication. The antenna capacitance is then set to 60 $\mu\mu$f, 100 $\mu\mu$f, 300 $\mu\mu$f, 500 $\mu\mu$f, and short circuit (infinite capacitance), and the alignment corrections are noted for each value of antenna capacitance. These measurements are usually made at 500 kc, 600 kc, 700 kc, 1000 kc, and 1400 kc for the broadcast band in cases where a high-inductance antenna primary is used. The 700-kc and 1000-kc points may be omitted if the alignment is good. If the antenna circuit employs a low-inductance primary, resonant above 1500 kc, the test points are usually 600 kc, 1000 kc, 1200 kc, 1400 kc, and 1500 kc. In this case the 1000-kc and 1200-kc points may be omitted if the alignment is good at 1400 kc. Corresponding values are used in the long-wave band if one is provided. In the case of receivers intended to operate with antennas having characteristics substantially different from normal, values of antenna capacitance appropriate to the case are chosen instead of the above-mentioned values.

ALTERNATIVE METHOD OF MEASURING ALIGNMENT OF PRESELECTOR CIRCUITS. To remove the possibility of error due to regeneration the avc line is biased by means of a battery, 4 1/2–9 volts will usually be sufficient. All circuits are then aligned in accordance with manufacturer's service instructions. At each test frequency the misalignment of the preselector circuits can be determined by means of a "tuning wand," which consists of a length of insulating rod, such as hard rubber or Bakelite, to which an iron-dust slug and a brass slug are attached, one at each end. This is held inside or near each preselector coil. If the output of the receiver decreases when either the brass or the iron is brought near the coil, no misalignment exists. However, if either end causes an increase in output the misalignment is equal to the decibel increase in sensitivity. If the increase is caused by the brass, the coil inductance is too high; if it is caused by the iron, the inductance is too low.

R-F STAGE MEASUREMENTS. The voltage gain from the grid of each r-f amplifier stage to the grid of the following stage is measured at all the regular test frequencies, and the 6-db band width is measured at three points in each band as discussed in the section covering antenna circuits. The output of the signal generator is fed into the grid of the tube incorporated in the stage being measured, and a suitable vacuum-tube voltmeter is connected to the lead which normally is connected to the following grid. When the signal generator is connected so as to replace a grid circuit which normally returns to the avc string it should be connected through a condenser, with a grid leak returning to the avc circuit so as to maintain normal bias on the tube. The grid leak is returned to the avc circuit even though the avc may eventually return to ground, since otherwise the measurements may be in error owing to the initial avc voltage caused by space current due to the emission velocity of electrons from the diode cathode.

I-F MEASUREMENTS. Diode Stage. The diode stage gain is measured by connecting the vacuum-tube voltmeter from the diode plate to ground, thus not appreciably disturbing the loading on the diode transformer. The signal generator output is connected through the previously mentioned condenser-grid leak combination, to the last i-f grid, and the stage is aligned for maximum output. A measurement of the voltage gain of the stage is then made, and the bandwidths 6 db and 20 db down are determined.

I-F AMPLIFIER STAGE GAIN AND BANDWIDTHS. Measurements of gain and bandwidths of i-f stages are made as described above except that the vacuum-tube voltmeter is not shunted across the grid circuit of the following tube; it replaces it. It often is advisable to isolate the vacuum-tube voltmeter by means of a high-resistance grid leak and condenser if these elements are not incorporated in the instrument.

MODULATOR STAGE GAIN AND BANDWIDTHS. This measurement is made in the manner described in the above paragraph. The signal-frequency circuits should be set to the middle of the band and the oscillator should be operating for this measurement.

CONVERSION GAIN. The connections for the signal generator and vacuum-tube voltmeter used in measuring the gain of the modulator stage are employed to measure conversion gain, the only difference being that the signal generator is set to the standard test frequencies in each band, conversion gain being defined as the ratio of rms voltage of intermediate frequency at the grid circuit terminals of the first i-f transformer to the rms voltage of signal frequency applied at the grid of the modulator tube. The avc system should be disabled for these measurements.

SENSITIVITY ON MODULATOR AND ON I-F GRIDS. The least signal input to the last i-f grid required to produce normal test output with the signal modulated 30 per cent at 400 cycles is called the i-f sensitivity. A corresponding measurement made from the modulator grid is called the modulator-grid sensitivity. The i-f circuits should be aligned in both cases.

DETECTOR SENSITIVITY. The detector sensitivity is defined as the amplitude of the 30 per cent modulated rms voltage which must be applied to the detector tube to produce normal test output. It is of course an indication of the total a-f gain. It is usually determined indirectly from the measured diode stage gain and the sensitivity on the last i-f grid, the difference between these two quantities being the detector sensitivity in decibels below 1 volt.

OSCILLATOR VOLTAGE. The oscillator voltage usually measured is the d-c voltage across the oscillator grid leak, as calculated from the measured resistance value of the grid leak and the measured current flowing through it when the oscillator is working. The voltage is measured at each test frequency throughout the range of the receiver. If the oscillator circuit necessitates some other method of measuring oscillator voltage, the procedure followed should be noted.

24. MISCELLANEOUS MEASUREMENTS ON A-M RECEIVERS

HUM VOLTAGE. The rms value of the hum in the output of the receiver can be determined by means of the output meter, provided sufficient sensitivity is available, or by connecting the primary of a voltage step-up transformer across the primary of the output transformer. The voltage measured across the secondary of this transformer by means of a vacuum-tube voltmeter is divided by the transformer voltage ratio to get the required value of hum in the output circuit. The voice coil should be connected for this measurement. The hum should be measured for both the maximum and the minimum settings of the volume control. To prevent receiver noise from interfering with the measurement a large condenser should be connected from the plate of the last i-f tube to the $B+$ line.

MINIMUM OUTPUT SIGNAL. This measurement is made to determine how close to zero output the volume control can reduce the signal. The receiver is tuned to a 1000-kc 46-db below 1 volt signal 30 per cent modulated, and the power across the standard dummy load is measured. The 400-cycle filter should be used to remove hum and noise.

CONDENSER GANG ALIGNMENT. In the case of gangs having uniform sections the corrections in micromicrofarads required to bring the antenna and/or r-f sections into alignment with the oscillator section are measured for several settings of the condenser gang. If these measurements exceed normal tolerances the gang can be knifed to make the necessary changes.

D-C VOLTAGES. The principal d-c potentials existing at the tube elements should be measured using a high-impedance voltmeter, preferably an electronic voltmeter, and tabulated. These measurements are made directly after the "as received" measurements and at standard line or test voltage.

POWER CONSUMPTION. The power consumed by the receiver at standard line or test voltage is measured by means of a suitable wattmeter. In the case of battery-operated receivers the A and B drains should be measured.

25. F-M RECEIVER MEASUREMENTS

Measurement of such characteristics as antenna and r-f gain and selectivity, i-f gain and selectivity, etc., is independent of the type of modulation which the receiver is designed to receive, hence the methods described above are applicable. In the case of f-m receivers, higher frequencies are involved; hence, the high-frequency measurement precautions to be described later must be observed.

EQUIPMENT. The equipment required for routine measurements on a-m receivers is usually required for f-m receivers, since relatively few are designed solely for frequency modulation. Generally, frequency modulation is one of several frequency bands incorporated in a receiver. Only measurements such as quieting and deviation sensitivity, etc., where f-m modulation of the signal generator is required, necessitate the use of an f-m signal generator.

DEFINITIONS OF TERMS. See also *Standards on Radio Receivers*, 1947, I.R.E., Methods of Testing Frequency Modulation Broadcast Receivers.

Standard Very-high-frequency Dummy Antenna. The very-high-frequency dummy antenna is a pair of resistors, one connected in series with each terminal of the signal generator, of such value that the total impedance between terminals, including the signal generator, is 300 ohms balanced to ground.

Standard De-emphasis Characteristic. The standard de-emphasis characteristic has a falling response with modulation frequency, the inverse of the standard pre-emphasis characteristic, equivalent to that provided by a simple circuit having a time constant of 75 microseconds. The characteristic may be obtained by taking the voltage across a condenser and resistor connected in parallel and fed with constant current. The resistance in ohms is equal to 0.000075 divided by the capacitance in farads. The standard de-emphasis characteristic is incorporated in the detector and/or audio circuits of the receiver.

Standard Test Frequencies. The standard test frequencies used in testing f-m receivers are 88, 98, and 108 megacycles. If more than three frequencies are required, it is suggested that 93 and 103 megacycles be included. If only one test frequency is needed, 98 megacycles should be used.

Standard Test Modulation. This term refers to a signal that is frequency modulated at 400 cycles with a deviation of 30 per cent of the maximum system deviation of 75 kc. The deviation due to standard test modulation is therefore 22 $^1/_2$ kc.

Maximum-sensitivity Test Input. The maximum-sensitivity test input is the least input signal of a specified carrier frequency having standard test modulation which, when applied through the very-high-frequency dummy antenna, results in standard test output when all controls are adjusted for greatest sensitivity. It may be expressed in terms of power in decibels below 1 watt, in decibels below 1 volt, or in microvolts. Generally it is advisable to use a 400-cycle filter to remove the noise from the output.

Maximum-deviation-sensitivity Test Input. The maximum-deviation-sensitivity test input is the least input signal of a specified carrier frequency having full rated system deviation which, when applied to the receiver through the very-high-frequency dummy antenna, results in 10 per cent distortion in the output when the volume control is adjusted to standard output. It is expressed in decibels below 1 watt, in decibels below 1 volt, or in microvolts.

Deviation-sensitivity Test Input. The deviation sensitivity test input is the minimum deviation at 400 cycles of a carrier of 60 db below 1 volt required to give maximum undistorted output when all controls are adjusted for greatest sensitivity. The deviation sensitivity is expressed in kilocycles or as a percentage of maximum full system deviation.

Quieting-signal-sensitivity Test Input. The quieting-signal-sensitivity test input is the least unmodulated signal input which, when applied to the receiver through the very-high-frequency dummy antenna, reduces the internal receiver noise to the point where the test output rises 30 db when standard test modulation is applied to the input signal, the volume control being adjusted to avoid audio overload. It is expressed in decibels below 1 watt, in decibels below 1 volt, or in microvolts.

26. OVERALL PERFORMANCE TESTS

The setup for overall performance tests is the same for f-m receivers as for a-m receivers, assuming that in each case an appropriate signal generator and standard dummy antenna provide the test signal. For measurements on f-m receivers a resistor having a value equal to the input impedance for which the receiver was designed may be used as a dummy antenna. If the receiver employs a balanced input, half the resistance is used in each side. The conditions set down on p. 11-43 regarding receiver alignment apply here also.

OVERALL SENSITIVITY. The sensitivity-test input for the maximum sensitivity, the maximum deviation sensitivity, the quieting-signal sensitivity, and the deviation sensitivity should be measured at 3 to 6 frequencies.

IMAGE AND I-F RATIOS. The procedure for determining these characteristics is the same as for a-m receivers. The maximum-sensitivity test input should be used in determining the image and i-f sensitivity.

OUTPUT AND AVC CHARACTERISTICS (LIMITER CHARACTERISTICS). This measurement corresponds to the output and avc characteristics of a-m receivers, and the same procedure is followed in obtaining the data. In the case of f-m receivers, the modulation percentages used are 30 per cent (22 $^1/_2$ kc) and 10 per cent (7.5 kc). The output power versus signal-input voltage is determined at 98 megacycles for maximum setting of the volume control. The limiter or avc characteristic is measured with 30 per cent modulation at a volume control setting which gives 6 db less than the maximum output. The output versus input for zero modulation is also determined.

HARMONIC DISTORTION. The measurements of harmonic distortion required for an f-m receiver are as follows: per cent harmonics versus per cent modulation and per cent harmonics versus signal-input level and per cent harmonics versus output.

OVERALL ELECTRICAL FIDELITY. The overall electrical fidelity is measured at 98 megacycles using a signal input 60 db below 1 volt, 30 per cent modulated. The curve thus obtained should normally be a close approximation to the standard de-emphasis curve for f-m receivers assuming no pre-emphasis in the signal generator. If the receiver is multi-band, the effects of the tone controls will be shown by the fidelity measurements on the a-m bands; otherwise, additional data should be taken to show these effects on the fidelity.

I-F AMPLIFIER CHARACTERISTICS. The overall selectivity of the i-f amplifier from the modulator grid to the first limiter grid or, if limiters are not employed, to the plate of the detector driver tube, is measured after the individual stage gain and selectivity have been measured. The level at the limiter grid should be that required for quieting the receiver. Where avc voltage is applied to i-f amplifier stages, the overall i-f amplifier selectivity characteristics should be measured for different levels of signal input to the modulator grid in order to show the detuning effect due to the change in tube input capacitance with a change in grid bias. Typical input levels are 80 db, 60 db, and 40 db below 1 volt. The avc voltage in the circuit should be held at the center frequency level by means of a battery supply.

27. SINGLE-STAGE MEASUREMENTS

REQUIRED TEST EQUIPMENT. Standard-signal generator, standard dummy antenna, output wattmeter, vacuum-tube voltmeter, tuning wand, high-resistance voltmeter, and wattmeter.

HIGH-FREQUENCY MEASUREMENT PRECAUTIONS. It will usually be found impracticable in making stage gain measurements on f-m receivers to substitute a vacuum-tube voltmeter for a tube or to place it in parallel with a tube. The inherent capacitance of the stage is usually so low that accurate measurements cannot be made if the measurement setup appreciably changes its value. A preferred method is as follows (Fig. 3): A resistor of approximately 500 ohms is substituted for the normal plate load of the tube which the vacuum-tube voltmeter would replace at lower frequencies. Under most conditions it is inadvisable to shunt this resistor across the normal plate load. The low-potential end of the resistor is bypassed directly to ground, keeping lead lengths as short as possible. The signal generator is then connected to the grid of the tube, the normal grid load being disconnected, and a vacuum-tube voltmeter is connected across the resistor in the plate circuit. The input-output voltage characteristic can then be checked and a suitable operating point chosen on the linear portion of the gain characteristic. With the suggested value of load resistance and the usual high-transconductance tube, a gain of 0–6 db will be secured.

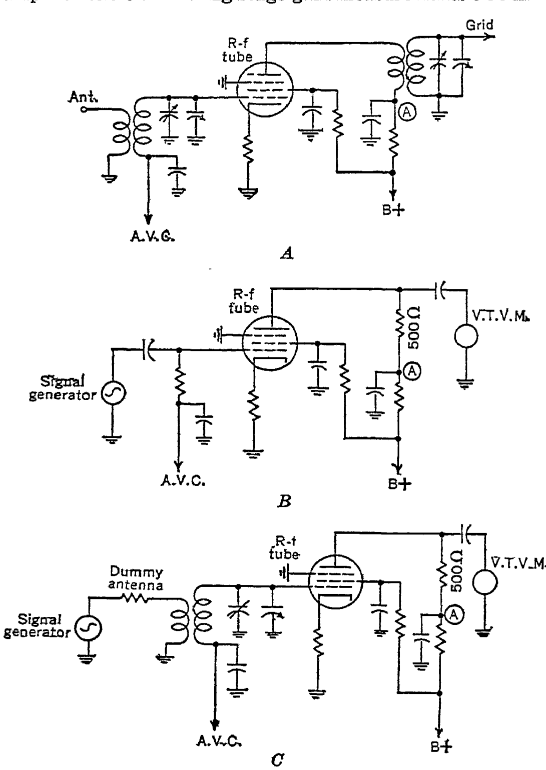


FIG. 3. Single-stage Measurement Set-up. *A*, normal circuit; *B*, calibrating rf tube; *C*, set-up for measurement of antenna gain and bandwidths.

Under these conditions it will be convenient to select as a reference output voltage that value which corresponds to a signal on the grid of 10 db below 1 volt. It is important that the voltmeter have sufficient sensitivity to make possible the selection of an operating point which can be obtained without danger of grid current due to high signal levels on the grid. After the operating point has been chosen, the grid of the tube can be connected back in the circuit, the signal generator connected to the grid of the preceding tube, and performance measurements made without upsetting the grid loading in the stage being measured.

ANTENNA GAIN AND BANDWIDTHS. The measurement of antenna gain and band width to provide data representative of those secured under actual operating conditions requires considerable care. If the first tube is a pentode r-f amplifier, antenna gain and band widths may be measured by placing the voltmeter across a resistive plate load of a few hundred ohms and then calibrating the tube as previously described. The data may then be secured without upsetting actual operating conditions.

If the antenna secondary feeds into the grid of the converter tube, it may not be possible to measure antenna gain separately. In such cases, it is usually possible to measure the gain from the antenna terminals to a resistive load in the plate circuit of the i-f amplifier tube. In this way the sum of the antenna gain and conversion gain can be obtained. If it is possible to measure the conversion gain separately, the antenna gain can be obtained indirectly by subtraction of the conversion gain.

R-F GAIN AND BANDWIDTH. The method of measurement of r-f gain, like that of the measurement of antenna gain, depends to a great extent on the type of circuit. In some cases, the voltmeter can be shunted across the grid of the modulator tube to enable a measurement of r-f gain to be made. In others, it is necessary to measure the r-f gain plus the conversion gain as has been discussed under the immediately preceding heading. It is always necessary to exercise care to make sure that the loading on the r-f tuned circuit under measurement conditions is the same as that under actual operating conditions.

CONVERSION GAIN. Conversion gain measurements where a pentagrid converter is used can generally be made by placing the vacuum-tube voltmeter across a resistive load in the i-f plate circuit as has previously been discussed for antenna and r-f gain measurements. Where a pentode modulator is used with both oscillator and signal voltages on the same grid, it is necessary to measure conversion gain indirectly. The antenna gain plus conversion gain or r-f gain plus conversion gain can be measured. The antenna gain or r-f gain, as the case may be, can then be measured and subtracted to give the conversion gain. As usual, in making measurements at these frequencies, care must be exercised so as not to allow the measurement setup to change the loading on the antenna coil.

I-F GAIN AND BANDWIDTH. The measurement of i-f gain and band width at the standard intermediate frequency of 10.7 megacycles can best be made by placing the vacuum-tube voltmeter across a resistive load in the plate circuit of the tube following the stage on which measurements are being made. The tube can then be calibrated and the measurements made without upsetting actual operating conditions.

REJECTION OF A-M BY F-M DETECTOR. No measurement has yet been standardized for this test. However, Fig. 4 shows one method for setting up equipment in a way which will provide useful information regarding the ability of an f-m detector to reject amplitude modulation. Imperfect a-m rejection at any signal input level will result in the "bow-tie" pattern. For this test the f-m signal generator must be capable of being simultaneously amplitude- and frequency-modulated. The a-m and f-m frequencies should be asynchronous. It is suggested that the f-m frequency be approximately 100 cycles and that the a-m frequency be approximately 400 cycles.

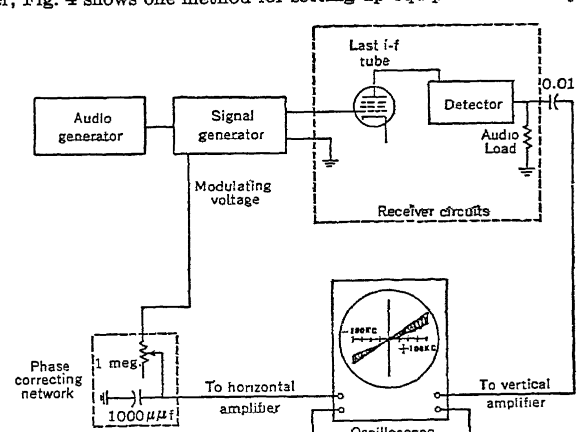


FIG. 4. Equipment Set-up for Test of A-m Rejection in F-m Detectors

OSCILLATOR DRIFT. The frequency drift of an f-m receiver oscillator circuit may be determined by employing the equipment setup shown in Fig. 5. The procedure is as follows:

1. Tune the receiver to the 100-megacycle signal from the crystal oscillator.

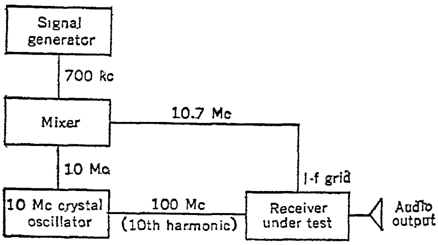


Fig. 5. Oscillator Drift Measurement Set-up

2. Adjust the frequency and level to produce zero beat in the output.
3. Maintain zero beat by adjusting the frequency of the signal generator. Note frequency drift versus time until the receiver has stabilized.

MISCELLANEOUS MEASUREMENTS. There are several characteristics for which measurement procedures have not yet been standardized. These include oscillator radiation, effect of detuning, two-signal interference tests, tests on built-in antennas, tests for the effects of downward modulation, and others. The measurement of such characteristics is important, but a lack of standardized procedures precludes the possibility of discussion at this time.

BIBLIOGRAPHY

I.R.E. Standards on Radio Receivers, Latest issue. *Methods of Testing F-M Receivers*, 1947. *Methods of Testing A-M Receivers*, 1948.
Foster, D. E., Receiver Characteristics, *Communications*, May 1939.
Characteristics of Broadcast Receivers, *The R.M.A. Engineer*, May 1940.
Matthews, A. C., Measurement of Receiver Characteristics, *Radio*, December 1944, March 1945.
Smith, F. Langford, *Radiotron Designers' Handbook*, Third Edition, Chapter 29, pp. 227–239.
Swinyard, W. O., Measurement of Loop Antenna Receivers, *Proc. I.R.E.*, Vol. 29, No. 7 (July 1941).

WAVE ANALYSIS

By E. Peterson

Some of the terms applied to devices which isolate, measure, and display the component amplitudes of complex electric waves include wave, harmonic, and spectrum analyzers; distortion and intermodulation meters and analyzers; and spectrographs.

28. WAVE CHARACTERISTICS

Waves to be analyzed may have components located at discrete frequencies or distributed more or less generally throughout frequency bands. Discrete components are encountered in studies of the response of a non-linear circuit or circuit element to the action of periodic waves. The resultant modulation products can be localized and measured individually or, in some cases, collectively. Band distributions on the other hand are found in speech and music and sometimes in noise. There, where the spectral distribution changes with time, it is difficult to keep track of individual components, and measurements are conveniently made of energies within bands of definite extent.

DISCRETE FREQUENCY DISTRIBUTIONS. Discrete or single frequency analysis is used in a variety of fields such as noise reduction, machinery noise, power interference, and modulation studies. Of these the last mentioned is the most important and the most demanding. Modulation studies are of interest throughout the frequency range occupied by the various communication services, and the amounts of power associated with them vary greatly in the different applications. Considering extreme cases, the power associated with modulation products in long loaded cables and in multichannel carrier amplifiers may be comparable with resistance noise in an audio band, while modulation products in high-power radio transmitters may amount to kilowatts.

The frequency of any modulation product is expressible in terms of the fundamental frequencies which produce it. If as example there are two fundamental frequencies f_1 and f_2, the frequencies of all products are included within the expression $|mf_1 \pm nf_2|$. Here m and n are positive integers, or zero. Products may be distinguished by their order, which is the sum of m and n. Thus the components $2f_1$, $2f_2$, $|f_1 \pm f_2|$ are all of the second order; $3f_1$, $3f_2$, $|2f_1 \pm f_2|$, $|f_1 \pm 2f_2|$ are all of the third order.

Modulation product amplitudes may be expressed in terms of the impressed waves or fundamentals producing the modulation, or in terms of a significant component, using arithmetic or logarithmic (decibel) scales for the ratios. Whereas in many detailed studies each modulation product of importance is specified individually, in others, in which an idea of the total modulation is required, the rms value of all unwanted products within a certain frequency band is specified. Thus, in one method of testing distortion in a radio link, a sine-wave signal, preferably at a low audio frequency, is impressed upon the transmitter. At the receiver output this component is removed by a filter or bridge type of suppression network. The rms value of all remaining products in the audio band is measured by means of a thermocouple or suitable rectifier. When expressed as a percentage of the rms fundamental in the normally terminated circuit, the ratio is termed the distortion factor.

When the modulation of a particular circuit element is under consideration, and is to be expressed in a way which would be applicable with the element connected in any given circuit, it is customary to specify the equivalent (modulation) generator emf. This is the voltage which, when inserted in the circuit made up of linear elements, gives rise to the observed distortion products. It is used ordinarily of elements only slightly non-linear.

The ratios of modulation products to their fundamentals which analyzers are used to evaluate fall into two ranges of values. One applies to products falling within the frequency range constituting quality impairment of the originating channel. The other characterizes products which fall outside the transmitted frequency range, constituting interference with other channels. Examples may be drawn from the audio, carrier, and radio broadcast fields. Thus the requirements on a system transmitting speech are usually that the distortion products be at least 20 to 40 db down on the wanted components in the normal channel (10 to 1 per cent in amplitude, or 1 to 0.01 per cent in power). Interfering products in other channels are restricted to smaller amplitudes. In multichannel carrier telephone amplifiers the figures range from 40 to 80 db down; in multichannel carrier grouping filters which serve to separate transmitted and received currents the figures range from 70 to 120 db down. In certain radio transmitters the products appearing in other channels are restricted to something of the order of 70 db down on the fundamentals. Corresponding figures for elements or units which make up a system may be more stringent than those specified for the system as a whole, especially where the modulations in several elements or units contribute to the total.

TEST WAVES. In the analysis of speech, music, or noise, or of the output waves of such devices as oscillators and harmonic producers, the questions involved are those of isolating and measuring the components. In establishing the performance of circuits or of circuit elements, however, the additional problem arises of specifying and of supplying the wave to be used for test purposes.

Generally speaking, the input wave should permit an indication of the performance of the apparatus to be tested, approximating as closely as possible the conditions of actual use. Where the apparatus is subjected in use to an impressed wave of simple form, comparatively little difficulty arises since the wave can be readily reproduced by oscillators and filter networks. On the other hand, the apparatus may be subjected to signal waves which are highly complex in nature, such as those of speech. Such waves may require comparatively elaborate and cumbersome setups for the measurement of modulation products, especially where the fundamentals and the various products overlap in the frequency spectrum.

To facilitate analysis, the complex signal waves are replaced by comparatively simple ones. At the present time one of two types is ordinarily used according to the problem encountered—a pure sine wave, or a complex wave consisting of two sine-wave components. To prevent interaction of the two components before reaching the circuit under test, tuned circuits, filter networks, or bridge networks may be employed. The amplitude of the testing wave is chosen so as to traverse much the same region of the non-linear characteristic as does the complex signal wave for which it is substituted. In some cases the test wave is made to have the same rms value, and in others the same peak value, as the normal signal. The test frequencies are so selected that, together with their important modulation products, they fall in a frequency region of interest.

To illustrate the use of single- and two-frequency test waves consider the investigation of a narrow band amplifier passing frequencies, let us say, from 100 to 110 kc. If a sine wave of frequency 105 kc is impressed for test purposes then the harmonics are 210 kc, 315 kc, and so on, all harmonics falling outside the transmitted band and being greatly attenuated. A measurement of distortion within the pass band would yield nothing. If, on the other hand, instead of applying a single frequency, we impressed two frequencies at 105 and 106 kc, say, the harmonics would fall outside the band and be attenuated as before, but the third-order products at 104 and 107 kc would lie within the band, as indicated in the spectrum (Fig. 1). These would furnish a useful indication of the non-linear distortion which a single-frequency test wave could not provide. Another example may be drawn from the same vacuum-tube amplifier. It may be shown that the amplitudes of the above-mentioned third-order products depend upon the impedance offered to second orders, one of which is the difference frequency (1 kc in the above example). This impedance may vary widely in different designs without affecting the normal transmission or

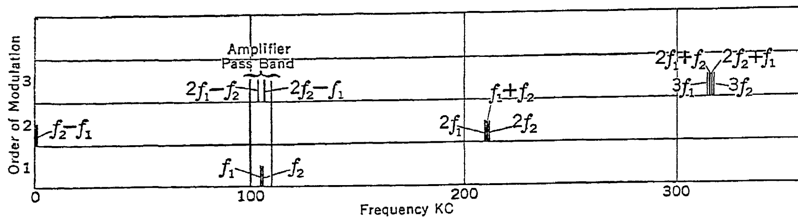


FIG. 1. Output Wave Spectrum of Narrow-band Amplifier with Frequencies of 105 and 106 kc Impressed

harmonic production. Figure 2 shows the results of measurement of third-order modulation as a function of the frequency difference between the two input components. No indication of this significant effect would be revealed by a single-frequency testing wave.

Another instance in which the two-frequency test wave is used is in the measurement of distortion produced by ferromagnetic-core coils. There only the main hysteresis loops are called into play when a sinusoidal magnetizing force is applied, but with complex waves of magnetizing force a different characteristic is involved since subsidiary loops appear.

To exemplify the use of the single-frequency test wave, consider a vacuum-tube amplifier supplied with power obtained from the 60-cycle line. There exists a certain amount of modulation of each signal component with the 60 cycles and its harmonics in the amplifier. To measure this effect a single-frequency test wave will serve.

BAND FREQUENCY DISTRIBUTIONS. Band frequency analysis involves measurement of energies or amplitudes in the sub-bands into which the main band is divided. It is usually employed for waves such as room noise or windage noise from certain types of machinery, which contain no prominent discrete frequency components, and for waves which contain discrete frequency components varying rapidly in both frequency and amplitude, such as speech and music. Band frequency analyses have also been used for waves consisting largely of discrete frequency components when it is unnecessary to know the precise frequencies of the individual components. Examples come up in studying the effect of vibration-absorbing mountings and of the acoustical treatment of rooms for the reduction of machinery noise. In the communications field, band frequency analyses of program material and of circuit noise are useful in determining requirements to be placed on systems for reproduction of the input waves with the desired fidelity and freedom from interference. For testing multichannel carrier systems, the loading effect of many talkers has been simulated by introducing bands of resistance noise.

FIG. 2. Third Order Modulation at Amplifier Output as Function of Frequency Difference between the Two Impressed Fundamentals

29. GENERAL ANALYZER REQUIREMENTS

The requirements which an analyzer is called upon to satisfy in general are those ordinarily imposed upon any piece of measuring apparatus: that it furnish an indication of the quantity sought, within certain limits of accuracy and of precision, without disturbing the essential performance of the circuit to which it is applied.

INPUT COUPLING. Connection of the analyzer to the circuit under test without altering the essential circuit performance is accomplished by assigning a suitable impedance to the input circuit of the analyzer. Three circuit arrangements have been used for this purpose. These are the high-impedance or voltage-analyzer connection, the low-impedance or current-analyzer connection, and the termination. In the first the analyzer is made to have an impedance much higher than that of the circuit across which it is shunted. In the second, the analyzer is made to have an impedance much smaller than that of the circuit in series with which it is connected. In the third, the analyzer is made to have a definite resistance to replace a resistance of the same value in the circuit studied. It will be observed that, under these conditions, insertion of the analyzer will normally have but little influence upon the normal functioning of the circuit studied, since all three may be made balanced or unbalanced to ground as desired. Figure 3 shows connections to a resonance type of analyzer as an example. For current analysis, connection is made to point a and for voltage analysis to point b when the resistor R is made large. For use as a termination R is given a suitable fixed value to fit the circuit or line to which it is connected, usually 600 or 72 ohms. The resistance r is usually made less than the effective resistance of the tuned circuit to avoid reducing selectivity and may be adjusted for the required gain.

A preferable connection for voltage analysis, shown in Fig. 3b, usually permits a greater fraction of the desired input voltage component to be transmitted while maintaining a high input impedance. Another connection suitable for voltage analysis, particularly useful over wide frequency bands, employs a probe including a condenser-resistance divider (Fig. 3c). Its input impedance is that of a 1- or 2-$\mu\mu$f capacitance shunted by a resistance of the order of 1 megohm. To make transmission flat with frequency, the two time constants C_1R_1 and C_2R_2 must be made equal.

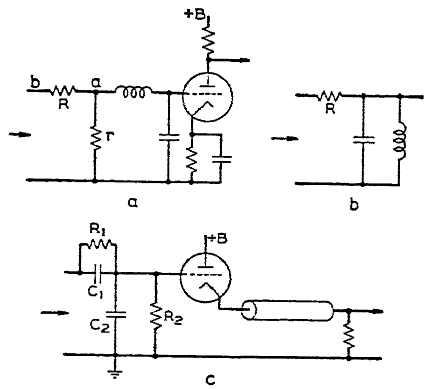


FIG. 3. Analyzer Input Circuits Illustrating High-input Impedance and Low-input Impedance Connections

RESOLUTION OF COMPONENTS. The analyzer band should be wide enough to permit easy tuning-in of the desired component, including an allowance for variations in frequency of the selected component. At the same time it must be narrow enough to select a specific product. The equivalent band width of a selective circuit may be defined as the band width of an idealized filter having a constant loss in the pass band equal to the minimum loss of the circuit in question, infinite loss outside the pass band, and passing the same amount of energy from a continuous constant-energy spectrum applied to the input. Analyzers for the study of low-frequency machine noise require a band width from 3 to 5 cycles. Those for telephone circuit noise caused by induction from neighboring power circuits require an equivalent band width of 20 to 30 cycles at least. For speech and for general types of noise, band widths of 45 to 300 cycles are appropriate. For intermodulation studies of discrete components, the bands used range from a cycle or so at low frequencies to several kilocycles and more at high frequencies.

Similarly the frequency discriminations required vary widely in different cases. In fixed-band analyzers used for speech, discriminations of 20 db or so against components located at the centers of neighboring bands are sufficient to provide useful information. Other a-f analyzers usually require discriminations of the order of 25 to 50 db against components 60 cycles from the tuned frequency.

The wide divergence in requirements has led to the development of a number of different forms of analyzers, each suited to a limited class of work.

MEASUREMENT AND DISPLAY OF SELECTED COMPONENTS. With the simpler manually operated forms of frequency analyzers, the point-by-point method of analysis is employed. Here, for the discrete frequency analysis of a steady wave, the frequency spectrum is explored point by point over the range of interest, and whenever a component of importance is located its frequency and amplitude are observed. For band frequency analysis of a periodic wave, measurements of the energy in each of a number of adjacent sub-bands are made successively without reference to the distribution within the sub-band. In analyzing waves of short duration which can be repeated a large number of times during the exploration of the frequency spectrum, a similar procedure is followed.

The net result of an analysis in which the several components are examined individually is usually presented as a meter deflection. To permit the use of simple d-c meters the selected wave feeds a thermocouple or a rectifier of the vacuum-tube or copper oxide type. The thermocouple is somewhat sluggish in operation, and the process of tuning accordingly must be slowed sufficiently to permit the products to produce an observable deflection. The slow response is advantageous where the smoothing out of rapid fluctuations in the selected wave is desired. Deflections are closely proportional to the square of the heater current. This is a valuable property in estimating energies in band analysis. Tube and varistor circuits may be made much more rapid in operation and may be given a wide range of desired relations between input amplitude and deflection by suitable choice of operating potentials and circuit parameters. For high rectifying gain a small negative bias may be used. For greater precision a large negative bias may be used since then a small input change may be made to yield a large change in rectified output. By the same token, however, an interfering component produces a correspondingly large change in rectified output.

Where a graphical record of wave analysis is to be had in a short period of time, one of several types of recording frequency analyzers is employed. These recording analyzers are designed to perform automatically the same operations performed manually with the simpler devices, including the plotting of wave component values. Recording analyzers of this type may be used as well for the analysis of non-periodic waves of short duration which can be repeated so that the effect of a steady wave is obtained.

Thus in the analysis of speech sounds and of single tones from musical instruments, or of other non-steady waves which are not readily reproducible, with devices employing the point-by-point method, the wave may be recorded and repeatedly reproduced. Finally, non-steady waves may be analyzed directly without the necessity of recording them by means of devices giving a visual indication of the complete energy spectrum of the wave and capable of showing the rapid changes in spectrum which are characteristic of speech sounds and some musical tones. A photograph of cathode-ray patterns or a recording on electrically sensitive paper of the indications of such an analyzer can be made for a permanent record of the analysis. These arrangements are discussed in article 31, p. 11-65.

30. METHODS OF WAVE ANALYSIS

The earliest methods used for wave analysis were based upon observed wave forms—observed directly by means of an oscillograph, or synthesized point by point from synchronous contactor measurements. Wave components are deducible from the wave form by application of Fourier series, which expresses the amplitude of any component as an integral. Evaluation of the integrals is customarily handled by a time-consuming point-by-point calculation, for which tables and schedules have been formulated to facilitate the work. This method is useful when only harmonics of a single frequency are present in the wave to be analyzed; when the wave has more than one fundamental, the computation ordinarily becomes too involved to be useful, since a multiple Fourier series must be employed. For harmonics of a single frequency which are not too small in amplitude, the accuracy of the determination depends largely upon the accuracy with which the oscillogram is read, the number of points included in the analysis, and the order of the harmonic. Because of its limitations, the computational method has been superseded by methods of direct measurement, wherever possible.

The first direct method of analysis was that of Pupin, in which a series-resonant circuit was used to select a component of interest, and amplitude was deduced from the indication of an electrostatic voltmeter connected across the condenser of the tuned circuit. Discrimination by means of simple tuned circuits is employed in certain types of analyzers of the present day, their sensitivity and flexibility being greatly increased by means of vacuum-tube amplifiers. Selectivity and sensitivity are improved by having several tuned circuit and amplifier units in cascade, and by substituting filter networks for the simple resonant circuit. A different way to improve selectivity involves the use of a negative resistance circuit to reduce tuned circuit losses.

RESONANCE ANALYZERS. Variable Tuned Circuit. The simple form of current analyzer shown in Fig. 4 is useful where a high degree of frequency selectivity is not required. It is made up of four units: coupling unit, frequency-discriminating network, amplifier, and indicator. The tuning unit precedes the amplifier to discriminate against other components of the input wave. This procedure reduces modulation in the amplifier, which may introduce components of the same frequencies as those to be measured.

The tuning unit consists of a simple series-resonant circuit with the input wave introduced through a small resistor. The amplifier input is taken across the condenser when the input fundamentals are of frequencies higher than the frequency of the component being measured. When the fundamentals or other possible interfering components are at lower frequencies the amplifier input is connected across the coil for greater discrimination. The amplifier is made responsive over the frequency range of interest; it is usually of the resistance-capacitance coupled type.

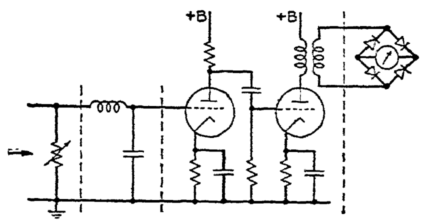


FIG. 4. Resonance Analyzer Circuit with Input Arranged for Current Analysis.

To increase the selectivity over that available in the simple tuned circuit, several discriminator-amplifier units are used in cascade. Efficient coupling between units is provided by step-down transformers. Here the signal level must be kept low enough to avoid appreciable modulation of the fundamentals in the early amplifier stages. A preferable arrangement from this standpoint would be to locate as much of the necessary frequency discrimination as possible ahead of the amplifiers. For greater discrimination than simple tuned circuits can conveniently afford, use is made of filter networks, which, being much more cumbersome and less flexible in adjustment, are usually employed where only a small number of fixed frequencies are to be measured.

The substitution method is currently used for evaluating component amplitudes since the loss of the tuned circuit, and therefore the analyzer response, varies with frequency. The connections are shown in Fig. 5, the switch being used to connect the analyzer to either the test or the standard circuits. The standard circuit includes an oscillator adjustable in frequency, a vacuum-tube voltmeter or thermocouple and associated meter to indicate the standard input, and a calibrated attenuator to vary the input amplitude to the analyzer. The analyzer input impedance is fixed at a value at which the attenuator is properly terminated.

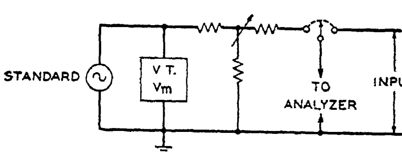


FIG. 5. Substitution Method for Evaluating Component Amplitudes

The test procedure consists in connecting the analyzer to the unknown, tuning in the desired product, and adjusting the analyzer gain until a suitable deflection is obtained on the output meter. The switch is then thrown to the standard side, and the standard oscillator frequency is varied until it coincides with the analyzer tuning. The attenuator is then varied until the same deflection is obtained as before. The frequency of the unknown component may then be found from the frequency calibration of the standard oscillator or of the analyzer, and the amplitude is computed from the voltmeter reading and attenuator setting.

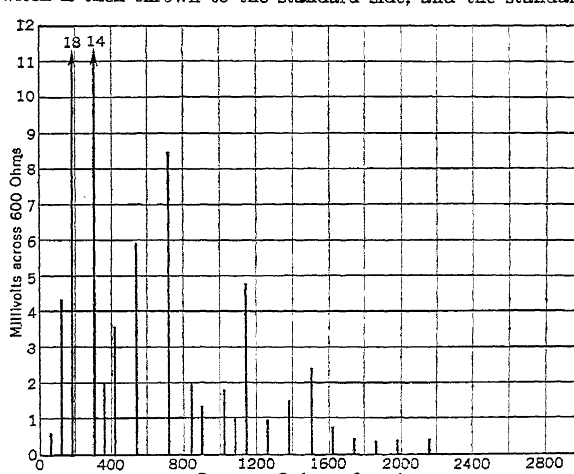


FIG. 6. Spectrum of Noise on an Open-wire Telephone Circuit

Another form of resonance analyzer with shunt tuning in the plate circuits of two amplifier stages has been used extensively in the analysis of power-frequency harmonics and of induction from power circuits. Voltages from 5×10^{-6} to 50 are measurable in the range up to 3 kc with an accuracy of ± 5 per cent. The discrimination of such an analyzer varies with frequency, being about 40 db at 60 cycles from the tuned frequency when tuned to 180 cycles, and from 24 to 32 db 60 cycles away when tuned to 3000 cycles. Figure 6 gives an example of an analysis of telephone-circuit noise caused by induction from power circuits. Approximately 1 hour is required to obtain with this apparatus the amount of data shown in the figure.

The speed and ease of operation of the more modern analyzers have resulted in supplanting the resonance type for all but special purposes.

Fixed Bands. A decidedly different form of analyzer using direct frequency selection avoids the need for manual tuning to each frequency of interest by means of a bank of fixed contiguous band-pass filters. Fourteen filters are used in one particular model to

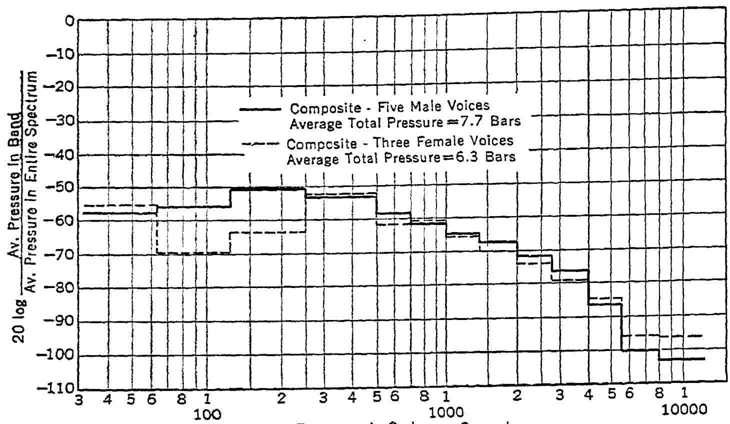


FIG. 7. Average Speech Pressures per Frequency Interval of 1 cycle per second—normal conversational voice. Distance 2″.

cover the frequency range from 30 to 12,000 cycles in approximately logarithmic steps. The filters are connected in parallel at the input side and introduce a loss of about 8 db at the crossover points and greater than 70 db over most of the frequency range beyond the transmission band. A flux meter is used to integrate the sub-band output over a 15-second interval for the measurement of average amplitudes, while a series of gas-filled discharge tubes with associated relays and electrical counters measures peak magnitudes over a 54-db range in steps of 6 db. These measurements are made in one band at a time although both average and peak measurements of the total spectrum may be made simultaneously with the band measurements. Peak measurements in $^1/_8$-second intervals are made over as long a time as desired. Figure 7 shows amplitude-frequency distributions for conversational speech for male and female voices.

Another type of fixed band analyzer has been used to measure distortion generated in non-linear circuits by complex input waves. Here the input wave is made to pass through a narrow band-elimination filter before it enters the circuit to be tested. A narrower band-pass filter is used for the analysis, with its pass band located within the band eliminated at the input. By this procedure the band-pass filter output has had the fundamental components eliminated, so that it constitutes a measure of distortion.

Feedback Type. A third means of providing frequency discrimination is particularly useful for the analysis of discrete components in sub-audio- and audio-frequency ranges. It consists of a negative feedback amplifier incorporating a bridge type of network in its feedback path. One form shown in Fig. 8 uses a parallel T network made up of capacitances and resistances which provides high suppression at a single frequency. At this frequency therefore the full amplifier gain is effective. At frequencies removed from the balance point, transmission through the feedback network acts degeneratively to reduce the amplifier gain. By insertion in the feedback path, therefore, the null of the network is transformed into a transmission maximum through the amplifier circuit. The network null may be conveniently varied in frequency by ganging the three resistors. Its selectivity remains constant on a percentage basis throughout the band, independent of frequency.

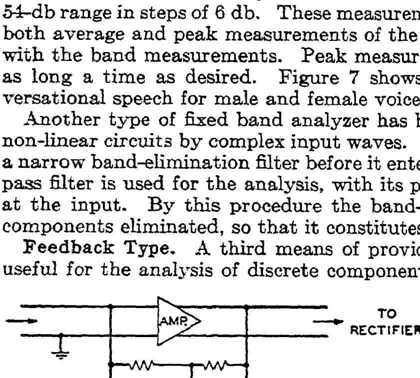


FIG. 8. Feedback Analyzer Using Twin-T Capacitance-resistance Network to Provide Frequency Discrimination

The following figures on frequency discrimination apply to a two-stage amplifier with a null set at 25 cycles. Roughly 12 per cent away from the null, the discrimination is 10 db; 20 per cent away it is 16 db, and in the limit approaches 45 db down on the maximum transmission.

Transmission of the amplifier proper is made substantially flat over the maximum range of analysis. Beyond that, the gain and phase around the feedback loop are arranged to avoid regeneration and oscillation, as in feedback amplifiers generally.

SUPPRESSION AND INTERMODULATION ANALYZERS. Although these two analyzer forms are not related in general, it will be convenient here to take them together.

A simple type of suppression analyzer previously mentioned evaluates the rms sum of the harmonics generated with usually a sinusoidal signal impressed. The fundamental may be suppressed through the action of a resonant bridge, or a suppression filter, or a twin-T network. The output includes any noise and interfering components which may be present. If the harmonics are comparable in amplitude to these components, the harmonic distortion proper cannot be found directly by this method.

Another form of suppression analyzer uses special procedures which are applicable to the detection of a carrier and its two sidebands, as practiced in radio broadcast reception. These procedures form the basis for a method of measuring intermodulation between two frequencies in the audio band, one low (40 to 100~) and the other comparatively high (1 to 12 kc). The two tones are supplied to the system under test, and the output is filtered to suppress the low-frequency tone. The residual wave is then treated as a carrier accompanied by the two sidebands to determine the extent to which the higher-frequency component is modulated by the lower, as described below.

Measurement of Percentage Modulation. The special procedures referred to may involve the cathode-ray oscillograph or the linear rectifier. If the modulating signal is applied to the horizontal deflection plates of the oscillograph, and the r-f output consisting of the carrier and its two second-order side frequencies is applied to the vertical deflection plates, then patterns like those shown in Figs. 9*a* and 9*b* are obtained on the oscillograph screen, according to the phase shifts in the system. These patterns yield the percentage modulation from the lengths of the minimum and maximum ordinates indicated. One hundred per cent modulation is obtained when the wave envelopes go to zero, and greater modulations are indicated by intersection of the envelopes. When a linear sweep circuit controlled by the modulating signal is connected to the horizontal deflection plates, rather than the modulating signal itself, then Fig. 9*c* is obtained.

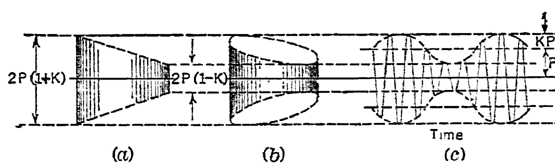


FIG. 9. Modulated Wave Patterns

Knowledge of the percentage modulation permits the second-order sideband amplitude ($KP/2$ in the figure) to be calculated accurately only when higher-order sidebands of the carrier are negligibly small. In that event the envelopes of Figs. 9*a*, 9*b*, and 9*c* are respectively linear, ellipsoidal, and sinusoidal. Otherwise the wave envelopes may be analyzed to yield the amplitudes of the second- and higher-order sidebands.

Another method for determining percentage modulation, generally used to evaluate intermodulation, employs a linear rectifier (envelope detector) to detect the modulated carrier wave with a minimum of audio distortion. The percentage modulation is then given as 100 times the ratio of the peak audio signal output to the d-c component. Where higher-order products are present, they may be evaluated by resonance or heterodyne-type analyzers.

HETERODYNE ANALYZERS. Dynamometer. Another method of analysis first used by Descoudres employs a dynamometer, in which the unknown wave and a standard wave of known frequency, amplitude, and phase are respectively connected to the two coils of the dynamometer. Since the dynamometer deflection is proportional to the product of the currents in the two coils, a constant deflection is obtained when the frequency of the sinusoidal standard is made equal to that of a component of the unknown wave. The magnitude and phase of the component are then found from a calibration when the standard phase is adjusted for maximum deflection. To determine the magnitude without regard to phase, the standard frequency is offset by a fraction of a cycle from that of the component under measurement, and the maximum swing is observed. The method is limited to comparatively low-frequency work.

From one standpoint the dynamometer may be regarded as fulfilling three functions: modulating the unknown with the standard wave, filtering the beat frequency output, and indicating the amplitude of the difference frequency component. Improvement in frequency response and in sensitivity has been obtained by replacing the dynamometer by a vacuum-tube modulator, and further increase in sensitivity is obtained again by amplifying the beat frequency output. As an indication of the component to be measured, use may be made of either the lower sideband or the upper sideband formed by the beating oscillator and the component under investigation. To select this product, mechanical, electrical, or piezoelectric filters and networks are conveniently located in the frequency range. These replacements of the functions discharged by the dynamometer result in the most widely used of all analyzers, the modern heterodyne type.

A particularly simple form of heterodyne analyzer has been adapted to the measurement of products not too small in relative and absolute amplitudes. In this arrangement no amplifier is used, and the mechanical movement of a meter connected in the plate circuit of a modulator tube serves to provide frequency discrimination as it does in the dynamometer.

General Tube Modulator Type. The heterodyne analyzer possesses a number of advantages over other types which have brought it into wide use. To mention the outstanding ones, first of all it uses a fixed discriminating circuit which is readily made highly selective and stable, and takes up little space. Next, the tuning-in of a desired band requires but a single adjustment—the frequency of the heterodyning oscillator. Finally the response can be made flat over an extended frequency range so that the substitution method of evaluating amplitudes is not required. Relative levels can be taken directly as the difference between attenuator settings.

Essential units of a representative model are shown in the schematic of Fig. 10. The principal elements comprise coupling unit, modulator, beating oscillator, filter, amplifier, and indicator.

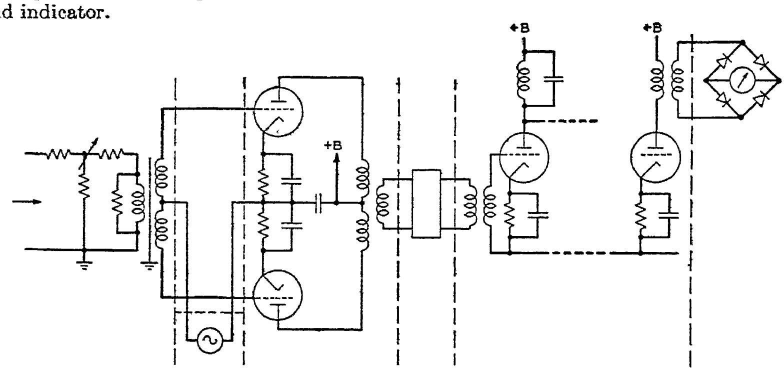


Fig. 10. Heterodyne Analyzer Circuit

The coupling unit is shown unbalanced to ground with a calibrated attenuator which serves to vary the input to the modulator. The analyzer therefore is used as a termination, but other coupling units for voltage or current analysis can be fitted to the input terminals of the attenuator.

The modulator is of the conjugate input type, to provide a convenient means for separating the signal and beat frequency circuits, to balance out the beat frequency oscillator wave in the output circuit, and to reduce the number of unwanted modulation components produced. Triodes have been used with the grid bias set close to plate-current cut-off. The shielded and balanced input coil shown should have uniform transmission over the frequency range of interest, and its modulation should be well below the amplitudes of the products measured. By the usual design procedure, the response of the modulator can be made uniform over a wide range of frequencies.

With values for noise, band width, and modulation which are readily attained, analyzers have been constructed capable of measuring second-order products 80 db down and third-order products 100 db down on their fundamentals. Here a three-stage amplifier was used with a gain of 130 db. The net analyzer gain including losses in the modulator and band filter was 120 db.

Where products larger in relation to their accompanying fundamentals are to be measured, the amplifier gain may be reduced. Where products smaller in relation to their fundamentals are to be measured, it becomes necessary to insert added discrimination against the fundamentals in the input to the modulator, so that we arrive at a combination of resonance and heterodyne types.

The oscillator used for heterodyning is designed to have a high degree of frequency and amplitude stability with supply voltage and temperature variations, and with time. Harmonics in the output are generally kept more than 40 db down on the fundamental. The oscillator amplitude applied to the modulator is made much greater than the signal input and slightly smaller than the grid bias, so that grid current does not flow in the modulator tubes. This amplitude, moreover, should be maintained fairly constant over the frequency range. All these requirements can be met readily with a bridge-type oscillator using thermistor or varistor stabilization. Frequency variation is most easily effected when the oscillator itself is of the heterodyne type, but frequency stability is not as high and spurious products have to be guarded against.

The ability of the analyzer to discriminate against neighboring input components is determined largely by the properties of the modulator tubes and by the selectivity of the selective network of Fig. 10, which may be made up of electromechanical, electrical, or piezoelectric elements. Which of these is used is a matter of convenience and of the frequency at which the band is to be located. Quartz crystals are generally used from 40 kc up to several megacycles. The band width is made narrow to cut down the noise introduced by the modulator to the greatest possible extent, as well as to reduce unwanted components. On the other hand, the band width should not be made so small that the tuning-in of the desired products becomes unnecessarily difficult and time-consuming. An equivalent band width of 5 to 20 cycles is found to be a useful compromise between the two requirements, the greater band width being used for measurements at radio broadcast frequencies. Analysis of waves in the short-wave region requires wider bands to accommodate incidental frequency variations; 2000-cycle bands have been used. The frequency at which the band is located depends upon the location of the frequency spectrum to be analyzed.

In order to avoid confusion of the measured components with extraneous products generated in the modulator, the band is placed either above or below the spectrum to be analyzed. Thus, in analyzers designed for use at carrier or radio frequencies, the band may be set at a low frequency; frequencies ranging from a fraction of a cycle to 1000 cycles have been used in various types. To avoid ambiguity in the measurement of small power-supply ripples or modulation products, the band should be made narrow and offset from harmonics of the power frequency. In an analyzer covering the range from 50 kc to 6 Mc, the crystal filter is located at 6.7 Mc and has a band width of 400 cycles. In audio analyzers the beat frequency is generally placed above the band; 11 kc, 50 kc, and 90 kc have been used. Figure 11 shows the frequency discrimination obtainable in several different types of selective circuits. The 100-cycle network referred to in that figure is a cumbersome multisection affair weighing close to 100 lb.

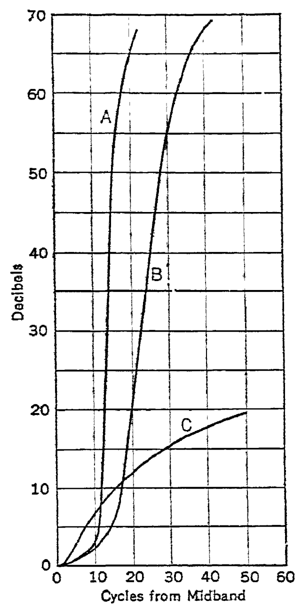


FIG. 11. Frequency Discrimination. *A*, electrical network, midband frequency, 100 cycles. *B*, piezoelectric network, midband, 50 kc. *C*, resonant circuit (Q = 100) tuned to 1 kc.

Since the amplifier is required to function over only the narrow band passed by the selective filter, it is conveniently built with screen-grid tubes or pentodes and tuned interstage coupling which permit of realizing comparatively high gains per stage. The tuned interstage coupling supplements the selectivity of the band filter and limits the noise output of the amplifier. An input transformer with a high step-up ratio may be used to minimize the effect of tube noise arising in the first amplifier stage.

Machine Noise Analyzer Unit; Mechanical Band-pass Filter. This device is used over the range from about 30 to 5000 cycles. It employs a mechanical band-pass filter working at 6000 cycles and having an equivalent band width of about 25 cycles. The effect of a double-balanced modulator is obtained in the electromagnetic structure through which the driving force is applied to the filter. In covering the above frequency range, the frequency of the heterodyning oscillator is varied from 6000 to 11,000 cycles. A discrimination of about 48 db 60 cycles from the tuned frequency is obtained in the four-section mechanical filter. The analyzer unit does not contain input or measuring circuits since it is intended to be used in conjunction with a noise meter which furnishes these circuits. It is readily portable. Figure 12 gives an example of the noise spectrum of a small synchronous motor obtained with this analyzer.

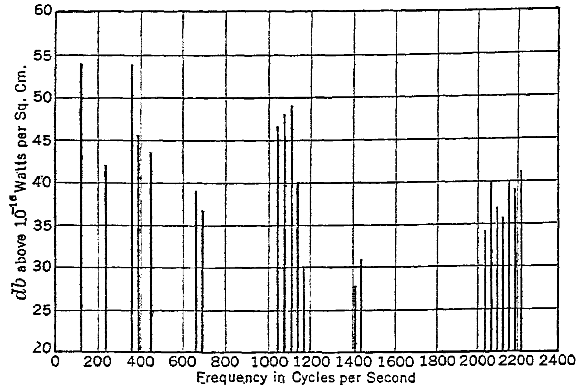


Fig. 12. Noise Analysis of Small Synchronous Motor—1/4 hp, 1800 rpm, 60 cycle, three phase, 220 volts, no load

General-purpose Analyzer Unit with Piezoelectric Band-pass Filters. With this device, it is possible to select either a 27-cycle band or 200-cycle band, from any part of the frequency range between 30 and 11,000 cycles. Lattice-type quartz crystal filters working at 50 kc are employed for both hands. For the narrow band, a discrimination of 52 db 60 cycles from the center of the band is obtained, while the wide band has a similar value 350 cycles from midband. The circuit includes a demodulator by means of which the wave components selected by the filter are translated to their original frequencies as indicated in the schematic of Fig. 13. This arrangement is particularly valuable in measuring closely spaced products, since it avoids output indication produced by transmission of the heterodyne source itself through modulator unbalance (carrier leak). The modulator leak here is translated to zero frequency and so is not transmitted by the a-c system connected to the demodulator output. The demodulator leak is far removed in frequency from the selected component and can be made harmless. Both modulator and demodulator are of the double-balanced copper oxide

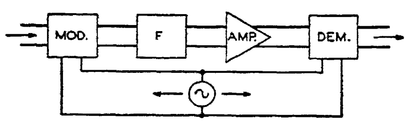


Fig. 13. Heterodyne Analyzer Circuit Used for Frequency Selection. The modulator-demodulator connection avoids spurious indications produced by leakage of the beating oscillator supply.

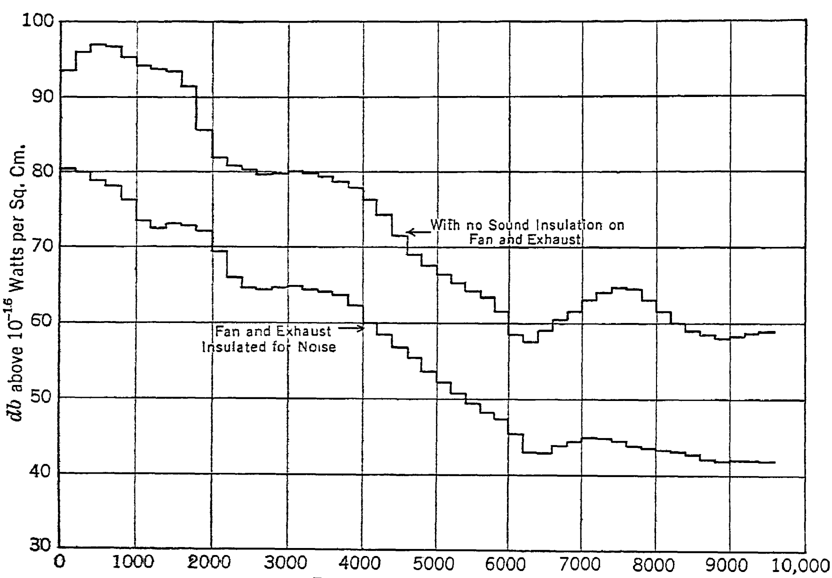


Fig. 14. Band-frequency Analysis of Exhaust Fan Noise

type and are supplied with carrier heterodyning power from the same oscillator through buffer amplifier stages. The analyzer unit does not include input or measuring circuits since it is intended as an attachment for a noise meter or similar measuring device of which these circuits are a necessary part. Figure 14 gives an example of the use of the 200-cycle band in the analysis of exhaust fan noise to determine the noise reduction obtained at various frequencies by means of sound insulation on the fan and exhaust.

31. SPECTROGRAPHS

Investigation of the spectrum of a complex wave by the methods just presented is a comparatively slow process at best since manual adjustment of a resonance frequency or of a beating oscillator frequency is required to shift the analyzer from point to point along the band to be studied. To speed up the process the analyzer may be tuned automatically, and the spectral distribution displayed in its entirety.

RECORDING TUNED CIRCUIT ANALYZER. Here the resonant frequency of a single electrical tuned circuit is varied in small steps over the frequency range by means of a player-piano pneumatic control. Simultaneously a photographic record is made of the current in the resonant circuit after amplification and rectification, the photographic paper being correspondingly stepped in position to accord with the frequency scale. Two frequency ranges are provided: 20–1250 cycles and 80–5000 cycles. The complete recording takes about 5 minutes. A 20-db amplitude range is the useful limit for a single record, the amplitude scale on the photographic record being approximately linear. Discrimination in the 80–5000 cycle band is about 20 db 60 cycles away from the tuning frequency, decreasing somewhat when the tuning frequency is above 1000 cycles. Figure 15 presents

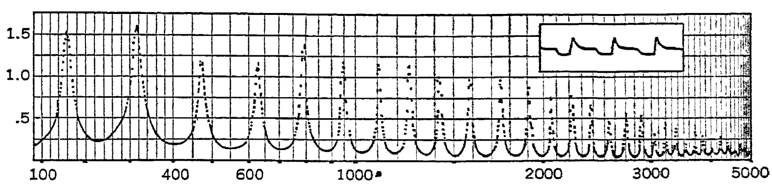


FIG. 15. Record of 160-cycle Buzzer Output

the analysis of a buzzer tone, each peak of the curve denoting the amplitude and frequency of an input component.

In any scheme of this kind which shifts a single discriminating circuit over the band, a definite compromise has to be made between the speed at which the circuit is shifted and the consequent resolving power, or ability to separate neighboring components. This relationship is discussed quantitatively below in the discussion of the Sweep Frequency Heterodyne.

TUNED-REED ANALYZER. This analyzer makes use of a series of electromagnetically driven reeds tuned to frequencies in the audio range, distributed with a uniform percentage difference in frequency. Each reed carries a small concave mirror which deflects a beam of light by an amount proportional to the oscillation amplitude of the reed. The vibrating

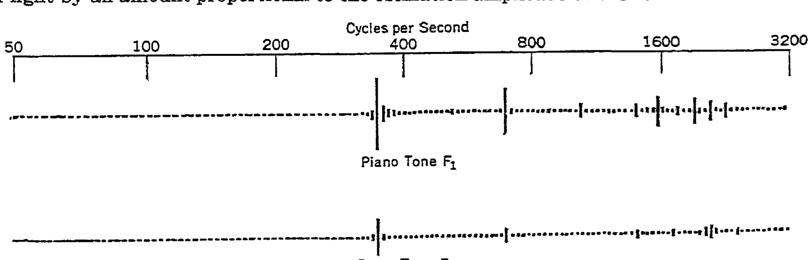


FIG. 16. Frequency Spectra Obtained with Reed Analyzer

spots of light indicate on a ground-glass screen the components present in the input wave which is applied to the electromagnets driving the reeds. Damping of the reeds is proportional to frequency and provides a discrimination of 20 db for components 2 1/2 per cent from the tuned frequency. A component 20 db down on the maximum can be evaluated with an accuracy of about 1 db when its frequency coincides with that of one of the reeds. A permanent record of a spectrum distribution constant in time can be obtained photographically. To record wave components varying with time a motion-picture camera must be used. Figure 16 shows spectrograms for two piano tones obtained with a demonstration analyzer of this type, covering the frequency range from 50 to 3200 cycles with 144 reeds.

COMMUTATED BAND ANALYZER. Selection of components is effected by a bank of contiguous fixed band filters extending over the frequency range to be covered. The output of each band filter is rectified, filtered, and amplified, as shown in Fig. 17. The vertically deflecting (*V*) plates of a cathode-ray tube are connected to each channel in turn by means of a commutator, while the horizontally deflecting (*H*) plates are connected to a synchronously operated stepped sweep. In this way, a linear frequency scale is laid out along a horizontal axis on the cathode-ray screen, and a vertical line represents the wave amplitude or energy within the corresponding band filter. The general aspect of spectrograms of this type is indicated on the oscilloscope screen depicted in the figure.

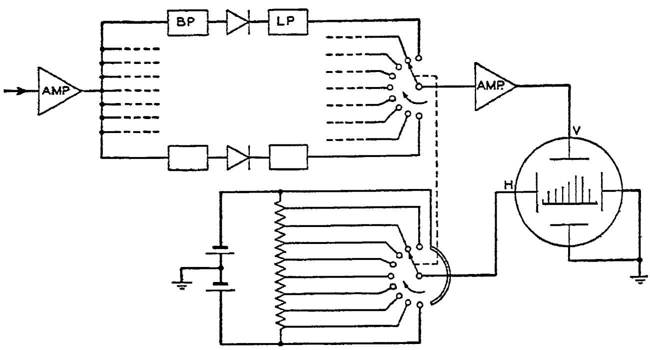


FIG. 17. Commutated Band Analyzer; Cathode-ray Presentation

From 10 to 30 filters have been used with band widths distributed linearly or logarithmically to cover the audio spectrum. With the linear distribution, bandwidths of 300 to 150 cycles have been used. Filter attenuations at the crossovers are 3 to 6 db, and at the mid-band frequencies of the immediately adjacent filters are of the order of 20 db. A non-linear compressor inserted in the lead to the vertically deflecting (*V*) plates of the cathode-ray tube helps to make the smaller amplitudes more readily perceptible. Accurate portrayal of an extended amplitude range, however, requires increase in filter selectivity to reduce interference from components falling within the attenuating regions of the filters. Commutation should be fast enough to reveal any envelope variations transmitted by the band filters. Thus if the widest filter band of the bank is B cycles wide, variations at a rate of $B/2$ are freely transmitted. The rectifier low-pass filter therefore should be $B/2$ cycles wide, or a little wider. For faithful indication, then, each channel should be sampled B times per second at least. While mechanical commutators have been shown for simplicity in Fig. 17, the channel sampling commutator may be replaced by electronic means, including clamps and a ring, with as many stages as there are bands to sample. Similarly the sweep commutator may be replaced by a properly synchronized electronic sweep developing a sawtoothed wave form.

SWEEP FREQUENCY HETERODYNE, CATHODE-RAY PRESENTATION. This method uses what is essentially an automatic heterodyne analyzer and is in general use throughout the radio frequency spectrum, from the broadcast band and below, to the centimeter region. A block diagram of the circuit arrangement is shown in Fig. 18. There a sweep generator provides a sawtooth wave at a sub-audio repetition rate for two purposes. First, it constitutes the horizontal sweep, and second, it provides a means for varying the beating oscillator frequency throughout a definite frequency range. To accomplish the second function, the sawtooth wave actuates a reactance tube associated with the frequency-determining circuit of the oscillator. Or, at ultra-high frequencies, the sawtooth wave is impressed directly upon an oscillator tube of the velocity-variation type. In either case, the oscillator frequency is made linearly proportional to the instantaneous amplitude of the sweep throughout its utilized portion.

The variable oscillator output then modulates the input wave to be analyzed so as to sweep its spectrum across the narrow pass band of the i-f amplifier. The i-f output is then detected, amplified, and applied to the *V* plates of a cathode-ray oscilloscope.

The amount of frequency variation (F) during a sweep cycle is made large enough to include the region to be analyzed. The sweep period (T) and the i-f band width (B) must be selected to accord with the resolving power required. If, for example, components equal in magnitude and separated by S cycles are to be displayed as distinct pips on the oscilloscope screen, the response produced by one of them must have built up and decayed to a sufficiently low value before the next component enters the i-f band. The transient response depends upon the i-f band width and has a duration of roughly $2/B$ seconds. This time is to be no greater than the time required for the sweep to traverse the frequency spacing between the two components, or ST/F. If then the i-f bandwidth is made, say, one-quarter the frequency separation between the components to be resolved, the sweep period T should be of the order of magnitude $F/2B^2$. Where the components to be resolved are not equal in magnitude, the sweep period or the sweep frequency deviation must be correspondingly reduced, if the two transients are not to overlap excessively. Resolution of a carrier and two sidebands is indicated on the oscilloscope screen of Fig. 18.

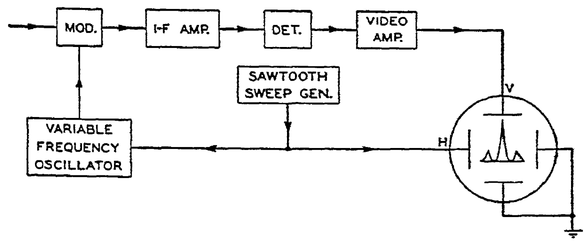


FIG. 18. Sweep-frequency Heterodyne Analyzer; Cathode-ray Presentation

In certain cases involving large numbers of components, it is desirable to display the envelope of the spectral distribution rather than to resolve the components individually. There the i-f band (B) is made wide enough to include a number of components, and the approximate relations given above remain applicable when the separation (S) refers to bands rather than to individual components. An example is shown in Fig. 19 of a spectrum envelope applying to a magnetron pulsed at 1800 cps, taken with an i-f band width of the order of 50 kc.

Various applications involve ranges of T from 1/100 to perhaps 1/2 second, of B from 2 to 100 kc, and of F from 100 kc to 200 Mc.

One of the most satisfactory methods for calibrating the frequency scale is to superpose on the modulator input of Fig. 18 a standard frequency source which is amplitude modulated by a sine wave at a known and adjustable rate. This superposes three pips at known frequencies on the signal spectrum, which can be varied in position to coincide with points of interest.

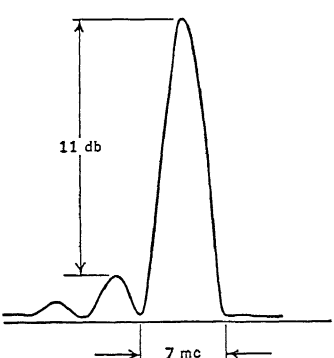


FIG. 19. Spectrogram of an Oscillating Magnetron Pulsed at 1800 ~; Pulse Duration Roughly 0.25 Microsecond. Wave components are not resolved.

For presentation of individual components a linear rectifier may be used to provide a linear output scale. For envelopes a square-law rectifier provides an output proportional to energy. Measurement of output levels may be carried out by one of two methods. The deflections may be read on a calibrated scale on the cathode-ray screen, or a calibrated attenuator in either input or i-f paths may be varied to bring to a fixed deflection the particular part of the distribution which is to be evaluated. In either method, overloading of the measuring system must be avoided.

A similar sweep type of heterodyne analyzer, developed primarily for the study of musical tones, has a galvanometer as output indicator A rotating mirror synchronized with the 10-cycle oscillator frequency sweep reflects the galvanometer light spot to a ground-glass screen. The rotating mirror thus provides the equivalent of a horizontal sweep. An electrical network is used at 20 kc with a band width sufficient to resolve tones separated by 200 cycles.

Another indicating and recording device used in conjunction with the sweep frequency heterodyne was developed primarily for noise and vibration studies in the audio band. It uses a scriber provided with a synchronous mechanical drive (like that of Fig. 20) which furnishes a permanent record of the spectral distribution upon waxed paper. Piezo filters make available analyzer band widths of 5, 50, and 200 cycles. The amplitude scale covers a range of 80 db. A complete record, directly legible, requires about 2 minutes' time.

SOUND SPECTROGRAPH. The sound spectrograph produces a visual record showing the distribution of energy within an audio band in both frequency and time. Though the development of this device must be regarded as still in the experimental stage, it constitutes as it stands a powerful means for the analysis of speech, music, and noise. Its power comes from the high concentration of information presented, which permits details of the spectral distribution to be followed as a function of time.

Figure 20 shows a simplified schematic of one form of the device which will serve to illustrate the basic idea. Three distinct functions are involved. First the sound to be analyzed is recorded so that it can be repeatedly reproduced. Here a magnetic tape recording is shown, mounted on a rotating disk which is driven by a synchronous motor. Second, analysis of the recorded sound is effected by a heterodyne type of analyzer in which the oscillator frequency is varied to move the analyzer filter in effect steadily over the sound spectrum. This variation is indicated in the figure by mechanical coupling between the varying condenser of the beating oscillator and the main drive shaft. In this way the analyzing frequency changes a small amount throughout each revolution of the shaft. Finally the analyzer output is recorded in synchronism with the reproduced sound. Recording is accomplished in the same manner as that practiced in facsimile reception. A drum bearing the recording paper is coupled to the main drive shaft, and the recording stylus is moved laterally a small distance in each revolution by means of a lead screw similarly coupled. The electrically sensitive paper mounted on the drum is marked by the stylus with gradations of density which accord with the analyzer output. The net result is to produce a sort of three-dimensional picture in which the energy distribution is depicted by density variations on a rectangular plot of frequency against time.

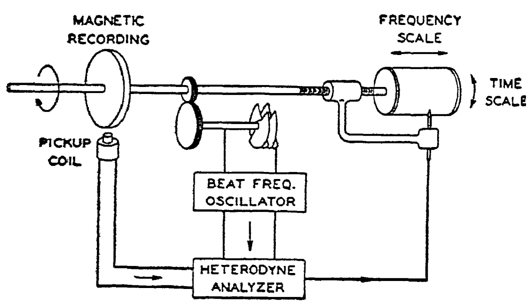


FIG. 20. Schematic Representation of One Form of Sound Spectrograph. The output of the sweep-frequency heterodyne analyzer is displayed as a pattern on electrically sensitive paper, exhibiting the energy-frequency distribution as a function of time.

A signal level range of 35 db is handled with the aid of automatic volume control which provides something like a threefold compression on a db basis. Two heterodyne band widths are available: one 45 and the other 300 cycles, as measured at the 3-db points. The linear frequency scale of the spectrogram is 2 in. high and covers the range from 100 to 3500 cycles. The linear time scale extends for a length of 12 in., corresponding to an original audio sample 2.4 seconds in duration. Two hundred rotations of the disk carrying the audio sample are completed in less than 5 minutes for the full analysis. The most recent development in this field speeds up the analysis by a factor of the order of 200.

Photographic reproductions of spectrograms of speech, music, and noise are to be found in the last four references listed under Spectrographs in the Bibliography. In the case of speech, the narrow band (45 cycles) analysis is adequate to resolve individual harmonics of the voiced sounds. The traces curve up and down as the pitch of the voice varies, and the spacing between harmonics gets bigger as the pitch increases at any particular instant. A definite loss of detail occurs when the wide band (300 cycles) is used since two or three harmonics are merged. A spectrogram of thermal noise shows energy concentrated in different frequency regions at different instants of time. Randomly spaced vertical spindles of the spectrogram correspond in length to the 300-cycle filter of the analyzer. Especially interesting are spectrograms of a warble tone which constitutes a frequency-modulated wave, produced by varying sinusoidally the frequency of an oscillator. The narrow filter reveals the presence of individual side frequencies of a tone warbled at a 50-cycle rate, but the wide filter cannot resolve them; they are integrated to reveal the instantaneous frequency.

The process of sound portrayal employed in the sound spectrograph described above results in a considerable simplification of apparatus where high resolution is desired. The equivalent machine to record high-resolution patterns directly would require something of the order of a hundred filters rather than a single one. Sound spectrographs of this type have been made to transcribe long samples of sound by employing the disk and drum arrangement of Fig. 20 with endless belts, one of magnetic tape and another of recording paper.

For special uses, such as experimental visual telephony for the deaf, low resolution patterns of the type provided by the sound spectrograph are formed at speech rates. Ten to twenty fixed filters or a scanning band affording equivalent resolution are used in this instrument, termed "visible speech translator." The outputs produce speech patterns in light upon a moving band of either phosphorescent or fluorescent material. Scanning and timing functions are controlled by synchronized electronic means. The accuracy and speed of these translators are sufficient to permit their use for the reading of continuous speech by trained observers.

BIBLIOGRAPHY

Resonance Analyzers

McCurdy and Blye, Electrical Wave Analyzers for Power and Telephone Systems, *Trans. A.I.E.E.*, Vol. 48, 1167 (1929).
Sivian, Dunn, and White, Absolute Amplitudes and Spectra of Certain Musical Instruments and Orchestras, *J. Acous. Soc. Am.*, Vol. 2, 330 (1931).
Scott, Analyzer for Sub-audible Frequencies, *J. Acous. Soc. Am.*, Vol. 13, 360 (1942).

Suppression and Intermodulation Analyzers

Wolff, A-c Bridge as a Harmonic Analyzer, *J. Optical Soc. Am.*, Vol. 15, 163 (1927).
Frayne and Scoville, Analysis and Measurement of Distortion in Variable Density Recording, *J. SMPE*, Vol. 32, 648 (1939).
Hilliard, Distortion Tests by the Intermodulation Method, *Proc. I.R.E.*, Vol. 29, 614 (1941).
Hayes, A New Type of Practical Distortion Meter, *Proc. I.R.E.*, Vol. 31, 112 (1943).
Pickering, Measuring Audio Intermodulation, *Electronic Industries*, June 1946.
Warren and Hewlett, Analysis of the Intermodulation Method of Distortion Analysis, *Proc. I.R.E.*, Vol. 36, No. 4, p. 457.

Heterodyne Analyzers

Cockcroft et al., An Electric Harmonic Analyzer, *J. I.E.E.*, Vol. 63, No. 337, 69 (1925); bibliography.
Landeen, Analyzer for Complex Electric Waves, *Bell Sys. Tech. J.*, Vol. 6, 230 (1927).
Moore and Curtis, An Analyzer for the Voice Frequency Range, *Bell Sys. Tech. J.*, Vol. 6, 217 (1927).
Suits, Thermionic Voltmeter Method for the Harmonic Analysis of Electrical Waves, *Proc. I.R.E.*, Vol. 18, No. 1, 178 (1930).
Arguimbau, Wave Analysis, *Gen. Radio Experimenter*, Vol. 8, Nos. 1 and 2, 12 (1932).
Castner, A General Purpose Frequency Analyzer, *Bell Labs. Rec.*, Vol. 13, 267 (1935).

Spectrographs

Wegel and Moore, An Electrical Frequency Analyzer, *Bell Sys. Tech. J.*, Vol. 3, 299 (1924).
Potter, Transmission Characteristics of a Short Wave Telephone Circuit, *Proc. I.R.E.*, Vol. 18, 583 (1930).
Hickman, An Acoustic Spectrometer, *J. Acous. Soc. Am.*, Vol. 6, 108 (1934).
Wolf and Sette, Some Applications of Modern Acoustic Apparatus, *J. Acous. Soc. Am.*, Vol. 6, 160 (1935).
Schuck, The Sound Prism, *Proc. I.R.E.*, Vol. 22, 1293 (1939); Panoramic Reception, *Electronics*, Vol. 14, 36 (1941); Recording Sound Analyzer, *Electronics*, Vol. 16, 100 (July 1943).
Williams, R.F. Spectrum Analyzers, *Proc. I.R.E.*, January 1946.
Gaffney, A Spectrum Analyzer for Microwave Pulsed Oscillators, *Waves and Electrons*, Vol. 1, 83 (1946).
Apker, Kahnke, Taft, and Watters, Wide Range Double Heterodyne Spectrum Analyzers, *Proc. I.R.E.*, 1947.
Montgomery, *Techniques of Microwave Measurements*, Rad. Lab. Series Vol. 11, Chapter 7, McGraw-Hill.
Koenig, Dunn, and Lacy, The Sound Spectrograph; Riesz and Schott, Cathode Ray Translator; Dudley and Gruenz, Visible Speech Translators with External Phosphors, *J. Acous. Soc. Am.*, July 1946.
Potter, Kopp, and Green, *Visible Speech*, Van Nostrand, 1946.
Kersta, Amplitude Cross-section Representation with the Sound Spectrograph, *J. Acous. Soc. Am.*, November 1948.
Mathes, Norwine, and Davis, The Cathode Ray Sound Spectroscope. *J. Acous. Soc. Am.*, September 1949.

General Surveys

Bourne, Wave Analysis, *Electronic Eng.*, Vol. 15, 149, 281, 472; Vol. 16, 31 (1942); 3, bibliography.
Scott, Measurement of Audio Distortion, *Communications*, Vol. 26, No. 4, 23 (1946).

MICROWAVE MEASUREMENTS

By E. W. Houghton

The electrical characteristics of any network can be completely expressed in terms of three quantities: impedance, power, and frequency. Attenuation ratio, although not a fundamental quantity, is of sufficient importance to warrant separate consideration. This section is concerned with methods and techniques for accurately measuring these quantities in the frequency range of roughly 1000–30,000 megacycles. The methods can be used to attain the accuracies summarized in Table 1.

Table 1. Microwave Measurement Accuracies

Quantity to Be Measured	Method	Accuracy
Impedance:		
Small standing wave ratios	Slotted line	±2%
	Directional coupler	±5% – ±0.2%
	Hybrid junction	±5% – ±0.2%
Large standing wave ratios	Slotted line	±5%
Power (averaged over several seconds)		
5–200 microwatts	Bolometer	±0.5 db
0.2–10 milliwatts	Bolometer	±0.1 db
0.01–100 watts	Bolometer-attenuator	±0.3 db
20–100 watts	Calorimeter	±0.3 db
Power (averaged over several microseconds)		
1–100 watts	Calibrated crystal	±0.6 db
0.1–100 kilowatts	Calibrated crystal	±1 db
Attenuation		
0–3 db	Bolometer	±0.05 db
0–13 db	Bolometer	±0.1 db
0–60 db	Heterodyne receiver	±0.2 db
0–40 db	R-f attenuator	±0.2 db
Frequency:		
Small differentials	Wavemeter	±0.005%
Absolute	Harmonic generator	±0.0001%
	Wavemeter	±0.01%

32. IMPEDANCE MEASUREMENTS

TRANSMISSION-LINE CALCULATIONS. At microwave frequencies practically all measurable impedances are in transmission lines (coaxial or wave guide), and the terminals of these impedances are transmission-line couplings (i.e., coaxial jacks and plugs or wave-guide choke or plane flanges) or simply specific transverse planes in the transmission line (see reference 17, p. 11-89). Absolute impedance values are rarely important, and in fact wave-guide transmission-line impedances can be defined in several ways (see reference 1). Ambiguity is removed by using a relative or "normalized" value which is the ratio of an impedance, Z, to Z_0, the characteristic impedance of the transmission line defined in the same way; thus $z = Z/Z_0$, where z is a complex value.

A section of uniform, lossless, transmission line, Fig. 1, transmits energy in a given mode between a source plane, s–s, and an impedance plane, t–t, by means of traveling electromagnetic waves. Unless the impedance at t–t terminates the line in its characteristic impedance, two such waves exist: (1) an incident wave which originates at the source and travels forward toward the impedance plane t–t, and (2) a reflected wave which originates at t–t and travels backward toward the source (see reference 2). The ratio of transverse voltage in the reflected wave to transverse voltage in the incident wave defines the reflection coefficient k, which is a complex quantity. Total transverse voltage and longitudinal current on the transmission line are the vector sums of the voltages and currents in the two oppositely traveling waves, which therefore combine to produce a standing-wave pattern distributed along the transmission line. The ratio of the maximum to the minimum total voltage defines (voltage) standing-wave ratio, S, which is not a complex quantity.

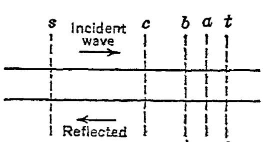


FIG. 1. Impedance Planes on a Transmission Line

The more general result of interference between oppositely traveling waves is to transform the impedance at t–t to new values at planes between s–s and t–t. The relationship between reflection coefficient and normalized impedance at t–t is: $K_t \angle\phi = (z_t - 1)/(z_t + 1)$, where z_t is a complex value. At some arbitrary plane a–a closer to the source and removed from t–t by line length $l/\lambda g$ wavelengths the reflection coefficient is $k_a = K_t \angle(\phi - 2\theta)$, where $\theta = 2\pi l/\lambda g$ radians. The normalized impedance at this plane is thus:

$$z_a = \frac{1 + k_a}{1 - k_a} = \frac{1 + K_t \angle(\phi - 2\theta)}{1 - K_t \angle(\phi - 2\theta)}$$

This can be expressed as $z_a = (z_t + j \tan \theta)/(1 + jz_t \tan \theta)$.

The oppositely traveling waves combine in phase opposition to give a minimum voltage standing wave at $2\sigma = (\pi - \phi)$ radians, or $(\pi - \phi)/4\pi$ wavelengths from t–t; at $\pi/2$ radians or $1/4$ wavelength closer to s–s the waves directly add to give a maximum voltage standing wave. The standing-wave ratio is thus $S = (1 + K_t)/(1 - K_t)$. From measurements of either K, or S, and σ, the normalized impedance at t–t can be computed (see reference 3) from

$$z_t = \frac{1 - j\,(S \tan \sigma)}{S - j\,(\tan \sigma)} \tag{1}$$

The transmission-line calculator (see reference 4), Fig. 2, provides a convenient graphical aid to visualization and computation of the above impedance transformations. An ex-

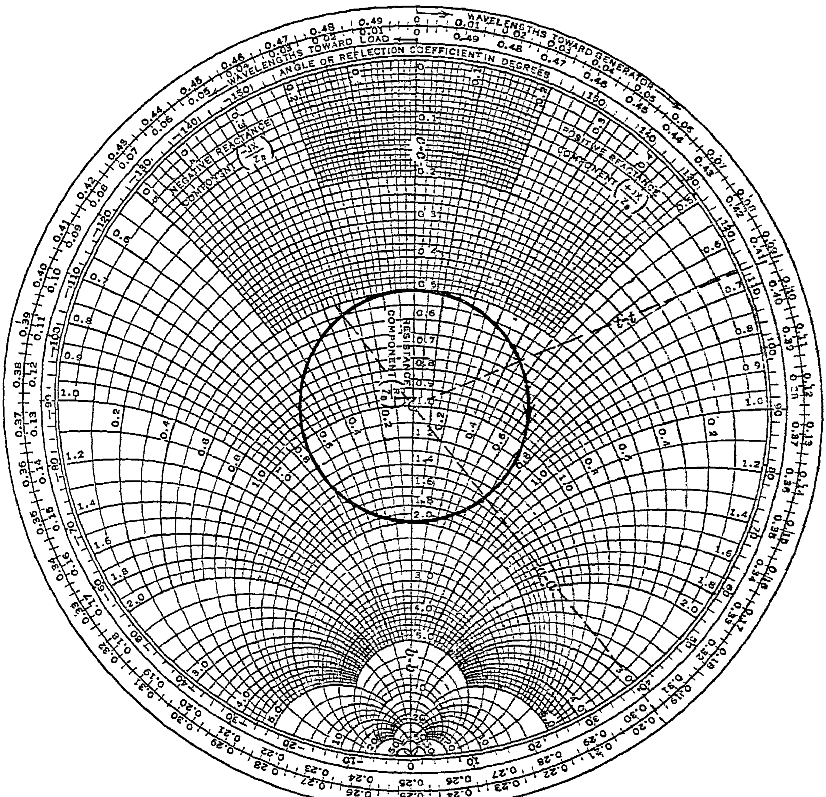


FIG. 2. Transmission Line Calculator

ample, drawn on the chart, will illustrate its use. As the line length between t–t and a–a increases, transformed impedance values are intercepted by the circle in the order indicated by the arrow on its periphery. Thus a normalized impedance $z_t = 0.65 + j0.45$ at t–t transforms to $z_a = 1.5 + j0.7$ at plane a–a located 0.102 wavelength from t–t. Normalized admittance values equivalent to series impedance values appear at points on the circle diametrically opposite; thus $y_a = 1/z_a = 0.55 - j0.25$. At b–b, z_t is transformed to a pure resistance $z_b = 2.0 + j0$; at c–c, z_t is transformed to $z_c = 0.5 + j0$, also a pure resistance. Planes b–b and c–c are the positions of the maximum and minimum standing wave voltages respectively, and the standing-wave ratio created by z_t is $S = z_b = 1/z_c = 2$. Following the above example in reverse order it is seen that the chart provides an extremely practical method for computing the impedance z_t from measurements which gave values of $S = 2$ and $\sigma = 0.408$. By this method eq. (1) can be solved rapidly and with a precision adequate for most measurements.

The results of impedance measurements are primarily used to study the electrical nature of the impedance, or to design networks for transforming this impedance to a new value (see reference 5) (usually the characteristic impedance of the transmission line, which it then terminates without reflection). However, in many cases a given impedance may be used to terminate an electrically long transmission line or a line of unspecified length. Under these conditions a measurement of the reflection coefficient phase angle would be trivial; thus in many practical cases impedance measurements are concerned only with ascertaining reflection coefficient amplitudes or standing-wave ratios produced by that impedance in a connected transmission line.

STANDING-WAVE DETECTORS. Of the three instruments most commonly used for impedance measurement the standing-wave detector is the most versatile since it affords information on both amplitude and phase of the reflection coefficient. It comprises: (1) a section of uniform transmission line (coaxial or wave-guide) which has a longitudinal slot in its outer conductor, and (2) a probe which projects through this slot and travels parallel to the axis of the line on a carriage. A standing-wave pattern of voltage is set up in the line when the test impedance is connected at one end and a source at the other. The probe, excited by this voltage, is connected to a detector whose output is a relative function of total voltage amplitudes along the slotted line. From readings obtained at the maximum and minimum voltage positions, the standing-wave ratio can be calculated, provided that the law of the detector is known or the detector has been calibrated. The detector may be a crystal (reference 6), bolometer (reference 8), or the mixer of a heterodyne receiver (reference 25). In the last case the mixer may be kept linear by operating at low levels; the second detector is used at a constant level by means of a calibrated i-f attenuator. The law of a bolometer detector and its bridge circuit can be obtained accurately by a low-frequency calibration. Specially selected crystals can be found to meet a square-law requirement over restricted input ranges, but they should preferably be calibrated against a bolometer power meter or heterodyne receiver.

A somewhat less accurate calibration method is to terminate the slotted line in a perfectly reflecting short and measure the standing-wave voltage distribution, which under ideal conditions approaches $V = C \sin (2\pi l/\lambda g)$, where C is a constant and $l/\lambda g$ is the displacement in wavelengths from a plane of zero voltage. A method for which the law need not be known is to operate the detector at a constant level by using a calibrated r-f attenuator between the slotted line and the source, or between the pick-up probe and the detector; standing-wave ratios are deduced from attenuator settings. The r-f attenuator method is satisfactory for measurement of high standing-wave ratios (10–100) but is generally considered inferior to other methods for low ratios (1 to 2).

In order to avoid reflection errors from the probe, its insertion is limited to small values (less than 10 per cent of wave-guide height or $^1/_4$ the difference between coaxial conductor diameters) for which its power output will be 20–30 db lower than the power in the slotted line. The most convenient oscillator sources are single-cavity klystron (reference 10) or Reflex (reference 8) oscillators which can be easily tuned over wide bands. After adequate padding these sources may deliver only 1–10 mw to the slotted line. High-sensitivity detecting and indicating methods are therefore often required. Three methods are common: (1) a heterodyne receiver can be used; (2) a simpler method is to 100 per cent modulate the source by a square wave (to avoid frequency modulation), then amplify and rectify the detector output in a high-gain amplifier-detector; (3) the simplest method is to use a sensitive galvanometer or microammeter driven directly by a crystal detector. A tunable transformer can be included in the probe-detector mechanism to give the highest sensitivity allowed by a given probe insertion; however, any variation in crystal or bolometer r-f impedance with standing-wave voltage variation may cause a change in the tuning and introduce a significant error. Since they are 15–20 db less sensitive than crystals, bolometer detectors are more satisfactory on high-power sources (magnetron test equipment, reference 11) for which high-sensitivity methods are not needed. For accurate measurements, stabilized power supplies must be employed, and spurious pick-up in microwave oscillators, high-gain amplifiers, and detector output leads must be avoided by complete shielding.

Design Requirements. The slot should be as narrow as possible, of uniform width, accurately centered in rectangular wave guide and parallel to the axis in the coaxial line, and long enough to allow a probe movement of at least $^1/_2$ wavelength. To prevent radiation the slot can be covered (by a contacting shoe or a shorting trap) for at least $^1/_4$ wavelength on each side of the probe. To avoid serious errors caused by small transverse variations of the probe in the slot the probe's outer conductor is sometimes imbedded in the center of a metal slug at least $^1/_2$ wavelength long, the bottom of which is flush with the inside of the outer conductor or wave-guide wall, and the sides of which definitely wipe (or definitely clear with very small gaps) the sides of the slot. In coaxial slotted lines the conductors must be accurately coaxial. The center conductor can be rigidly supported and centered at the source end, but at the load end it should be supported (if necessary) by a very thin washer of low-dielectric-constant material; the residual reflection (pure shunt susceptance) can be calibrated and allowed for subsequently. An assortment of low-reflection fittings (tapers, jacks, plugs, adapters) may be required for connection to the load, although greatest accuracy can be attained with as few attachments as possible. Wave-guide slotted lines are preferably terminated with a well-surfaced plane flange, to which adapters may be attached, or with a standard choke on one end and choke-cover flange on the other so that either type of connection is available by reversal of the slotted section.

The most serious requirement is that the bottom of the probe shall travel accurately parallel to the axis of the transmission line. The probe carriage may travel on the outside of the line itself, or on ways rigidly attached and made accurately parallel with the axis. It is generally possible by accurate machining or electroforming techniques (reference 12) to maintain parallelism within ± 0.0002 in. for at least $^1/_2$ wavelength of travel.

Accuracy. On standing-wave ratios above about 2, inaccurate assumptions (or calibrations) for the law of the detector and the amplifying-indicating system, and meter-reading errors, can easily limit the measurement accuracy to 10 per cent; by using a calibrated r-f attenuator, or preferably a heterodyne receiver, or by deducing the voltage standing-wave ratio from measurements confined to the region of the voltage minimum (reference 3), the accuracy can be improved to nearly that for low standing-wave ratios. Below ratios of 2, percentage accuracy is limited primarily by: (1) change in characteristic impedance and end reflections introduced by the slot; (2) reflections introduced by load connectors (flanges, jacks, tapers, or adapters), unless these are to be considered a part of the unknown impedance; (3) reflections from the probe; and (4) non-parallelism of probe travel. The effect of (1) can be deduced by measuring an impedance through two lines of the same dimensions, one with and the other without a slot. Sometimes errors from (2) can be reduced by determining an equivalent shunt susceptance (reference 16), but this is impractical when the connector introduces multiple discontinuities. It is possible, but not very easy, to measure and calculate errors from (3) (reference 13); this effect can be reduced by maintaining a matched impedance looking toward the source and eliminated by withdrawing the probe until there is no significant change in power delivered to the load (as monitored by another detector); but a compromise must usually be effected between this error and the error from (4) since (4) is reduced by larger probe insertions.

Experience has shown that carefully constructed coaxial standing-wave detectors can be used with an accuracy of about 5 per cent and the accuracy for wave-guide detectors is usually about 2 per cent.

DIRECTIONAL COUPLERS. Directional couplers (reference 14) are commonly included as a permanent section of the coaxial or wave-guide transmission line in microwave systems for monitoring incident and reflected power (reference 15). In such applications they (1) introduce negligible reflection in the transmission line (S less than 1.05), (2) create almost no high-power arcing problem, and (3) contain no moving parts. Compared to standing-wave detectors, directional couplers can be used for routine measurements of reflections with somewhat less accuracy by standard techniques, and with much higher accuracy by special techniques; they can be more conveniently used (1) to tune a load impedance to match the line, (2) to tune a source impedance to match the line, and (3) to correct for changes in interaction loss between the source and load impedance.

Power levels proportional to the power in the reflected and incident waves can be measured at the input and output ends respectively of an auxiliary transmission line coupled to the main line by means of two equal-sized probes, loops, or orifices separated by approximately $^1/_4$ transmission-line wavelength (reference 14). Illustrated in Fig. 3, this instrument comprises a directional coupler in its simplest form. A directional coupler can thus be used to measure reflection-coefficient amplitude, which is the ratio of the square root of the two measured power levels. Bolometer power detectors are convenient for high-power sources, and on low-power sources the same sensitive detecting and indicating methods previously described are applicable. It is preferable to switch the same detector-indicator alternately between incident and reflected wave outputs (alternately terminating the other output) so that only a ratio calibration is required. Low-reflection wave-guide and coaxial switches are convenient for this purpose, or special arrangements can be employed to transmit both waves to the same output. Detector law uncertainties can be eliminated by using a calibrated r-f attenuator to equalize incident and reflected powers. The attenuator method is best used for measuring nearly matched impedances, since, for example, a change in standing-wave ratio from 1.01 to 1.02 changes

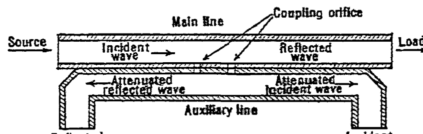
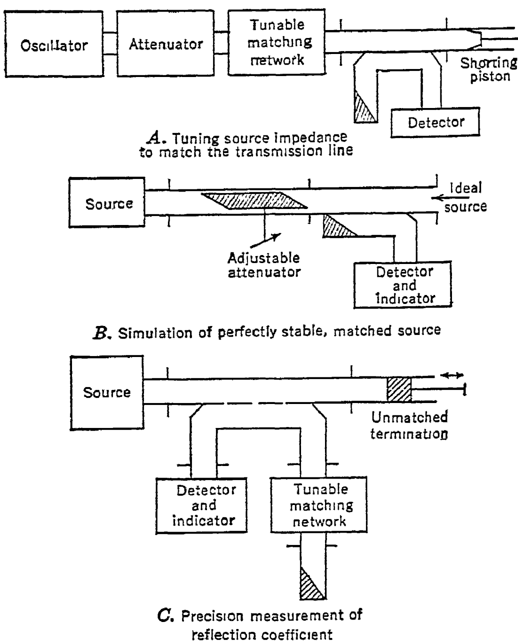


FIG. 3. Directional Coupler in Waveguide

the reflected wave output power by 6 db whereas one from 2.02 to 2.04 changes the output power by only 0.12 db.

A directional coupler can thus be used as a sensitive indicator for tuning a load to match the transmission line, and, because reflection coefficient is continuously monitored, it is especially convenient.

In Fig. 4 a directional coupler is used for matching a source impedance to the line. As a shorting piston (or preferably a reflector with a constant K of 30 to 50 per cent) is moved back and forth in the line, the tuning controls are adjusted until the incident wave output remains constant. The oscillator must be sufficiently masked or decoupled to prevent changes in its frequency and efficiency as the load impedance varies.



A. Tuning source impedance to match the transmission line

B. Simulation of perfectly stable, matched source

C. Precision measurement of reflection coefficient

FIG. 4. Directional Coupler Applications

Without tuning for it, the condition prevailing for a matched source impedance (Fig. 7) can be simulated by monitoring and keeping the incident power constant by means of a variable attenuator ahead of the coupler, Fig. 4. The variation of power in the load as its impedance changes is reduced to that of reflection loss only. Power-time instability in the oscillator can also be corrected for by this method.

Design Requirements. Directional couplers for impedance measurement should not have losses much below 20 db (see p. 11-72) to avoid serious loading by the coupling holes and interaction between them. By careful machining or electroforming techniques (reference 12), inside dimensions and the wall thickness of the main line must be made precisely uniform over the longitudinal region occupied by the coupling slits or holes, which must be constructed with exacting tolerances. Transverse slits have been used both in coaxial lines and in wave guides (usually in the wide side), but for wave-guide couplers round holes in the narrow side are usually better since they give higher losses which can be more accurately controlled by standard precision boring techniques. In coaxial lines the center conductor must be accurately centered; reflection from any support at the load end cannot be calibrated out of the measurements which ordinarily do not contain phase information. Similarly, careful attention must be given to designing reflectionless fittings and connectors.

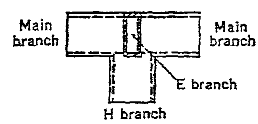
The most serious requirement for accurate measurements is that the directional coupler be designed for the frequency band in which it is to be used. Its directional coupling action (directivity) is frequency sensitive. As λg, the transmission-line wavelength (not to be confused with free-space wavelength) varies, the effect of a fixed hole spacing is to add a spurious vector voltage to the true reflected wave output in the auxiliary line. The ratio of this spurious voltage to the incident wave voltage (also in the auxiliary line) shall be designated "unbalance reflection coefficient," K_u. In the worst case, a measured value of reflection coefficient will be in error by $\pm K_u$. For a simple two-hole coupler in wave guide, K_u is zero at a λg approximately 1 per cent greater than four times the hole spacing, but for a ± 0.6 per cent, ± 1.3 per cent, and ± 3.2 per cent change in λg the values for K_u are approximately 0.01, 0.02, and 0.05 respectively. The band width can be greatly increased by increasing the total number of holes, and the simplest pattern is to separate equal-sized pairs of holes (spaced $1/4$ the mid-wavelength) by $1/2$ mid-wavelength. By this method K_u can be kept lower than 0.015 over an 8 per cent wavelength band in a four-hole wave-guide coupler and over a 20 per cent band for an eight-hole wave-guide coupler (reference 26).

Another spurious vector, adding to the true reflected wave output, comes from partial reflection of the incident wave from an imperfect termination of the auxiliary line. Where it is practical to position this termination for first a minimum then a maximum reading, it is possible to nearly eliminate its error by averaging the two readings.

Accuracy. In the laboratory where there is usually no serious limitation on length, it is practical to use four- or eight-hole couplers containing broad-band terminations with *VSWR* less than 1.05 in coaxial and 1.02 in wave guide. Under these conditions, routine measurements can be made with a *VSWR* accuracy of about 10 per cent in coaxial and 5 per cent in wave guide, since detecting-indicating systems can be made to contribute small percentage errors. In microwave systems directional couplers are often limited in size so that the realizable accuracy is usually lower.

Figure 4 shows a method wherein the auxiliary line termination is tuned to cancel the K_u of the holes, giving the effect of a perfect coupler and termination. As a mismatch of constant K is moved back and forth in the main line, the termination tuning is adjusted until the reflected wave output remains constant (not zero). Alternatively, if a termination of $K = 0$ is used the output can be directly tuned to zero. Loads (including their connectors) connected to the directional coupler can then be measured with an accuracy of about 0.1 to 0.2 per cent. Accuracy is limited primarily by the detecting-indicating equipment.

HYBRID JUNCTIONS. Hybrid junctions (reference 18) can be used in many of the applications described for directional couplers, compared to which they may possess certain advantages in sensitivity, size, design simplicity, and band width. It is possible to make hybrid junctions in either wave-guide or coaxial lines. However, a design which will give accurate measurements over a wide bandwidth does not yet exist for coaxial lines. Descriptions and examples in the following discussion are applied specifically to wave-guide hybrid junctions for which satisfactory designs are commonplace.



A. Hybrid junction in waveguide

B. Measurement of reflection coefficient

FIG. 5. Hybrid Junction

A type of hybrid junction commonly used for impedance measurement is formed by joining an H-plane T (off the narrow side) and an E-plane T (off the wide side) to a straight section of rectangular wave guide (reference 12). Projections of the central axes of both T's meet at a common point on the axis of the main guide, Fig. 5. This geometrical symmetry makes the E and H branches conjugate. When the main branches are terminated by equal impedances there is no transmission between E and H branches. Hybrid junctions can therefore be used to tune a load on one main branch (test branch) to match that on the other main branch (reference branch); the degree to which the tuned load matches the characteristic impedance of the wave guide depends upon (1) the intrinsic balance in the junction and (2) the reflection coefficient of the reference termination. The directional coupler has corresponding limitations; in fact, both devices separate out direct and reflected waves by a cancellation process. However, it is sometimes more helpful to visualize the hybrid junction as the microwave equivalent of a low-frequency hybrid transformer (reference 18) with the added complication that the input impedances at the junctions, unless matched, are transformed to new values along the transmission-line branches.

Looking toward the junction the impedances are inherently unmatched to the branch transmission lines; on well-constructed junctions the *VSWR*'s looking alternately into E, H, test, and references branches are approximately 2, 3, 1.3, and 1.3 respectively when the other three branch lines are terminated in Z_0.

By methods similar to those described for the directional coupler, hybrid junctions can also be used for routine and precise measurements of reflection coefficient. Figure 5 illustrates the circuit configuration. The E-branch output (detector input) is proportional to the vector sum of the reflected waves in the test and reference branches, and a wave created by junction unbalance. When the latter two are negligible, the detector input is proportional only to the power reflected from the unknown impedance; reflection coefficient can be deduced from the ratio of this power to the power incident upon the unknown. For sensitivity calibration a power level proportional to the incident wave (in the test branch) is most conveniently applied to the detector by substituting a shorting plunger for the unknown.

As a consequence of its reflection coefficient, k_t, the test branch carries a *total* reflected wave, V_r, which is not uniquely proportional to the product of the reference incident wave V_i and the unknown reflection coefficient k_x. In fact, $V_r = V_i k_x/(1 - k_x k_t)$, which is a complex quantity. Evidently the amplitude ratio V_r/V_i as measured on the hybrid junction is not the true reflection coefficient K_x; therefore the maximum error in routine measurements created by K_t (assuming that the effect is not canceled by purposely phasing the unknown along the line) is $1/(1 - K_x K_t)$. For example, a measurement which gave $K_x = 9.1$ per cent ($SWR = 1.2$) may be in error by a maximum factor of 1.01 since $K_t = 13$ per cent. The error increases at higher values of K_x, but nearer zero the error from this source will be negligible (reference 27).

When a short circuit is substituted for the unknown for sensitivity calibration, $K_x = 1$. As the position of the short circuit is varied the power in the detector will deviate from the reference value by maximum factors of $1/(1 + K_t)$ to $1/(1 - K_t)$. From an observation of the maximum and minimum detector outputs it is possible to deduce that reference detector output which is proportional to the reference incident power in the test branch. However, the impedance presented to the source will be seriously altered by the short circuit in the test branch, so the above factors can be applied exactly only when the oscillator is well masked and the source impedance is well matched to the line; alternatively the source's incident power can be monitored by a directional coupler and kept constant, Fig. 4. The detector must also present a well-matched, Z_0, load. An alternative method of calibration which reduces the interaction effects occasioned by introducing a shorting piston is to substitute a known impedance (calibrated by other means) of medium or low K. In this case, the net measurement accuracy can be no better than that of the calibrated impedance.

The above interaction effects can be materially reduced or even eliminated by introducing properly positioned susceptances (references 5 and 17) (matching posts, windows, or dielectric blocks) which cancel the junction's discontinuity reactances and transform the impedance of the H and E branches so that each branch is matched when the other three are terminated with matched loads. Under this condition, the main branches are also conjugate, and power sent into any one branch divides equally between the two nonconjugate branches (reference 18). Such an "ideal" hybrid junction can be used in all those applications described for directional couplers, except monitor, without disturbing the transmission line, with as good accuracy and higher sensitivity. For example, high- or low-impedance mismatches can be measured accurately, a load can be tuned, and a source impedance can be tuned to Z_0 by monitoring and tuning for constancy the incident wave output on one main branch while moving a shorting piston in the other main branch.

The above impedances may be matched over only a relatively narrow band, unless the posts or windows are put right in the junction. However, the unbalance reflection coefficient may be seriously increased by this method of impedance matching. Completely satisfactory wide-band solutions have not yet been found. Matching networks are not essential for many measurement requirements.

Accuracy. Accuracy of routine measurements is primarily limited by (1) interaction between unknown and test branch impedances, (2) reflection coefficient of the reference termination, and (3) junction unbalance reflection coefficient. For low mismatches (K less than 5 per cent) the first error is negligible, and under this condition the same techniques described for directional couplers can be used to (*a*) cancel the second error by positioning the reference termination and (*b*) cancel both second and third errors by tuning the reference terminations for precision measurements. On hybrid junctions constructed by precision electroforming techniques (reference 12) the unbalance reflection coefficient can be kept less than 0.005 over the entire wave-guide pass band. Therefore on low mismatches the accuracy of routine measurements can be as high as or higher than on well-constructed directional couplers with the added advantage that lower-sensitivity, more stable detecting and indicating equipment can be used.

Compared to directional couplers, hybrid junctions have the following advantages: (1) they are simpler to design for a low unbalance reflection coefficient over a wide band (geometrical symmetry is the only requirement); (2) they are smaller; and (3) they are more sensitive (the reflected wave power in the detector is at least 10 db higher than that for a 20-db directional coupler). However, (1) they cannot be incorporated in transmission lines (with negligible reaction) to monitor reflection coefficient; (2) interaction between "unknown" and test branch impedances may introduce a significant error in measuring reflection coefficients higher than about 5 per cent ($S = 1.1$); and (3) impedance interaction effects make it more difficult to calibrate sensitivity without introducing a calibration error.

33. ABSOLUTE POWER MEASUREMENTS

LOW-POWER MEASUREMENTS BY BOLOMETRIC METHODS. Microwave powers less than 10–20 mw are measured by bolometric methods which are, so far, the only ones available for accurate measurements in this frequency and power range. A transmission line carrying the unknown power is terminated in a bolometer detector, the only absorbing element of which is a thermally sensitive resistor. Its d-c resistance changes when r-f power is dissipated in it, and ideally the change is independent of the frequency of excitation. The resistance change can therefore be related to r-f power by a low-frequency or d-c calibration. Alternatively the resistance can be biased to a given value and kept constant when r-f power is applied by removing an equal and measured quantity of d-c or low-frequency power.

A bolometer detector (Fig. 6 is typical) comprises a thermal resistor and a reactance network for matching it to the transmission line. Sensitive thermal resistors are small, essentially "lumped," elements. Two types are commonly employed: thin, short, filament wires (reference 19), and bead thermistors, the thermal element of which is a tiny bead made up of a mixture of metallic oxides (reference 8) (see Table 2). The r-f resistance of the thermal resistor may be different from d-c resistance, but, in order to compare r-f power directly against d-c (or low-frequency) standards, (1) its d-c resistance change must be dependent only upon incremental heating power (when external temperature is constant), (2) the heat distribution from r-f power must be equivalent to that generated by a uniformly distributed d-c current, and (3) all the r-f power absorbed by the detector must be dissipated in the thermal resistor.

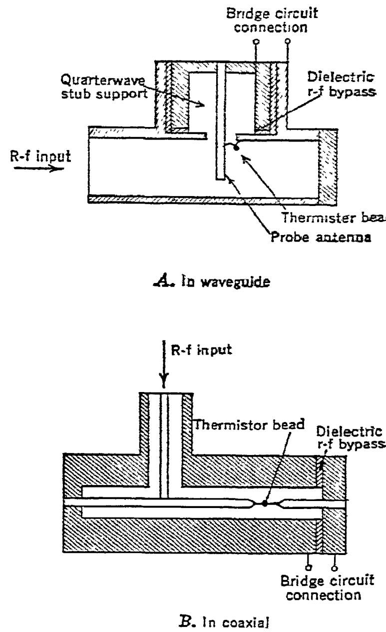


FIG. 6. Pre-tuned, Wide-band Bolometer Detectors

Requirement 1 is met on all thermal resistors when space-integrated values of resistance and heat are specified. However, the time constant of the thermal resistor may be of such a value that, in two extreme cases, resistance changes either (*a*) exactly follow or (*b*) completely fail to follow the modulation envelope of the r-f power (when amplitude modulated). In either case the average resistance change is proportional to average power, but in (*a*) envelope peak power can be measured. As sinusoidal modulation frequency is increased from zero (all other parameters remaining constant) the resistance modulation will decrease to half its maximum value at 36 cps and at 450 cps, respectively, for the thermistors and platinum wire listed in Table 2.

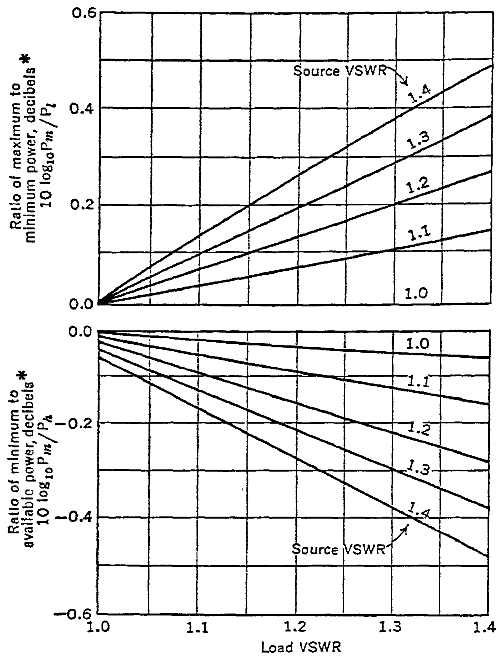
Table 2. Sensitive Thermal Resistors—Typical Characteristics

	Platinum Wire	Thermistor Bead		
Temperature	25° C	25° C	25° C	25° C
Resistance	200 ohms	200 ohms	125 ohms	50 ohms
Biasing power	15.3 mw	10 mw	13.5 mw	23.5 mw
Resistance-power coefficient	+4.5 ohms/mw	−29 ohms/mw	−14 ohms/mw	−4.8 ohms/mw
Power-temperature coefficient *	−0.05mw/deg cent	−0.1 mw/deg cent		
Time constant †	350 μsec	2500 μsec		
Safe maximum current	11 ma	200 ma		
Dimensions	0.118 × 0.00006 in.	Lead wires, 0.001 in. diameter; Bead, 0.020 × 0.010 in.		

* For constant resistance; also the ratio of power to temperature changes which produce the same resistance changes.

† For 67 per cent of ultimate resistance change.

Requirement 2 is met on filament wires by (*a*) designing a matching configuration which places the r-f current maximum at the midpoint of the wire, and (*b*) by limiting the upper frequency to that corresponding to a free-space wavelength of about 8 times the length of the wire (reference 7). Platinum wires have given satisfactory results up to 10,000 megacycles. Bead thermistors, which are more nearly ideal "lumped" elements, have been used up to 25,000 megacycles, where they have shown discrepancies less than 5 per cent when checked against calorimeter standards.



*For the above VSWRs, line lengths are chosen so that P_l = least power and P_m = most power in the load. Available power P_a could be delivered if conjugate impedence-matching transformers were used in the line.

FIG. 7. Power Delivered by a Mismatched Source to a Mismatched Load

Requirement 3 must be met by transforming the a-c resistance component to the transmission line's characteristic impedance with a matching network in which considerable care is taken to eliminate extraneous circuit losses from sliding contacts, joints, and conductors. This problem is simplified by (*a*) biasing the thermal resistor to a reasonable resistance (50–200 ohms) and by (*b*) designing a low-*Q* matching configuration, which tends to eliminate high current concentrations (standing waves) in the detector.

The power absorbed (and therefore measured) by a detector load is a function of both source and load impedances, and, depending upon relative phase angles (or line length), varies between the limits shown in Fig. 7 (reference 20). To eliminate these uncertainties in measurement and interpretation of power quantities it is evidently desirable to use a detector that presents a nearly matched impedance. Wide-band, pretuned detectors, in addition to being especially convenient to use, inherently meet the low-*Q* requirement. Table 3 lists the band widths that have been realized for typical pretuned bolometer detectors using thermistors.

Since the r-f impedance varies with it, the thermal resistor's d-c resistance must be kept nearly constant by varying (with temperature) the biasing power by methods peculiar to the bridge measuring circuit.

Table 3. Bandwidth-Impedance Characteristics of Typical Pretuned Thermistor Detectors

Frequency Band in Megacycles	Maximum VSWR	Transmission Line	Thermistor Operating Resistance in Ohms
5 to 600	1.2	Coaxial	Characteristic *Z* of line
700 to 1500	1.4	Coaxial	100
2900 ± 17.2%	1.4	Coaxial	100
3700 ± 8.1%	1.4	Coaxial	100
4500 ± 11.1%	1.4	Coaxial	125
4100 ± 12.2%	1.2	Wave guide 2 in. × 1 in. OD	75
4600 ± 8.2%	1.1	Wave guide 2 in. × 1 in. OD	75
9.050 ± 6.1%	1.4	Wave guide 1 1/4 in. × 5/8 in. OD	125
24,000 ± 4.2%	1.4	Wave guide 1/2 in. × 1/4 in. OD	125

Bolometer Bridge Circuits. Biasing and power-indicating circuits are designed to accommodate characteristics such as those shown in Table 2. Resistance changes are most accurately detected in bridge circuits such as Fig. 8 which illustrates the simplest method of introducing biasing power and measuring r-f power. Power can be directly measured as the difference of the two d-c biasing powers required for reference resistance, one before and one after r-f power is applied. However, this simple method contains three important disadvantages: (1) since the r-f power may be a small difference of two relatively large d-c powers, small meter-reading errors may introduce large errors in the difference power measurement; (2) r-f power is not continuously indicated; (3) the unbalance caused by external temperature changes subsequent to initial balance conditions are indistinguishable from r-f power changes.

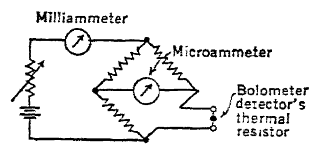


FIG. 8. Simple Bridge Circuit for Bolometer Detectors

Difference errors, (1), can be reduced by interposing a resistance network between the bridge and a current or voltage source which is kept constant at a precisely measurable single value. The network must contain accurately known resistance elements, one or more of which are switched in or out in small and known steps. Biasing powers are then deduced from settings on the switch.

Within the accuracy limitations imposed by (3), unbalance current can be used as a continuous indication of relative r-f power level; absolute power can be deduced from a calibration of the unbalance sensitivity. This sensitivity calibration will be an uncritical but not negligible function of temperature unless the d-c voltage across the bridge can be kept constant (this can be made possible by using variable low-frequency biasing power to balance the bridge initially). Resistance unbalance must be limited to a maximum value dictated by the detector's r-f matching requirements.

The lowest power that can be accurately measured is limited by (3) and the temperature coefficient of the thermal resistor. For example, from Table 2, a subsequent temperature change as small as ± 0.1 deg cent would cause a 3-db error in measuring powers of 2.5–5 μw and 0.1 db in measuring powers of 0.25–0.5 mw. These errors can be materially reduced by (*a*) using thermal insulation, (*b*) using large metal masses to limit the rate of temperature change, (*c*) interposing between the source and the detector an essentially instantly operable cut-out switch, reactive gate, or attenuator which can then be used to remove r-f power rapidly to check the initial balance, and (*d*) using temperature-compensating bridges. A combination of all the above techniques is usually required for accurate power measurements in the range of 5–100 μw.

A practical solution to the disadvantages of the simple bridge circuit has been to combine all the remedies into a more complex circuit such as Fig. 9. The thermal resistor in

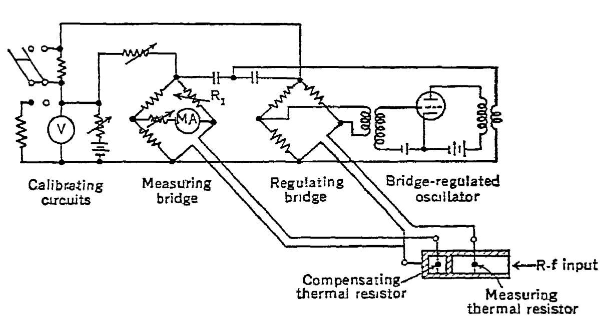


FIG. 9. Temperature-compensating, Self-calibrating, Direct-reading Bridge Circuit for Bolometer Detectors

the bolometer detector is used as the regulating element in a conventional bridge-regulated oscillator. The oscillator automatically delivers enough power to the thermal resistor to bias its resistance very close to the bridge-balancing value (within 1 per cent). The oscillation power level will therefore vary with ambient temperature with a coefficient determined by the temperature coefficient of the thermal resistor. From Table 2, this coefficient would be -0.05 mw/deg cent for the platinum wire and -0.1 mw/deg cent for bead thermistors.

The output of the low-frequency oscillator is also delivered to another bridge which contains a compensating and measuring thermal resistor (never excited by r-f power) of the same type as that used in the bolometer detector. When R_1 is properly adjusted the power received by the compensating resistor varies with temperature at a rate equal to that required by this resistor to be biased to a constant resistance. Both thermal resistors must be subjected to identical external temperature variations. Intimate thermal coupling can best be accomplished by burying the compensating thermal resistor in the same metal mass (preferably large) which forms the outside boundaries of the bolometer detector.

As a net result the measuring bridge remains balanced, even though temperature changes, until r-f or incremental d-c power is applied to the bolometer detector. The d-c biasing powers are kept constant so that the unbalance sensitivity of the measuring bridge does not vary with temperature. The overall sensitivity is established at one power level by using the d-c calibrating network which introduces a known increment of d-c heating power in the bolometer detector. Other unbalance current readings can be related to the calibrated point from a knowledge of the indicator law. Both the measuring bridge network and the intrinsic characteristics of its thermal resistor influence the law of unbalance current versus power input. In general, this law is not exactly linear, but it can be accurately calibrated by exciting the detector with a convenient low-frequency source whose relative output power is varied in known ratios by a low-frequency attenuator.

Accuracy. It is usually practical to limit errors from r-f sources to very low values, and then power-measurement accuracy is primarily governed by the d-c calibrating and indicating circuits. The circuit of Fig. 9 can be used to measure full-scale powers with less than ± 0.1 db error, half-scale powers within ± 0.2 db, and quarter-scale powers within ± 0.3 db. On an indicating meter whose resistance is equal to the bridge arms, deflections of 200 and 50 microamperes per milliwatt are typical for thermistors and platinum wires respectively. Therefore powers below $^1/_4$ mw cannot be very accurately measured unless meter-reading errors are kept low by using sensitive galvanometers or very stable indicator amplifiers; it is then possible to attain accuracies of ± 0.4 to 0.6 db down to 5–10 μw. However, the severe temperature drift problems must be solved by exacting methods as discussed above.

MEDIUM AND HIGH-POWER MEASUREMENTS BY BOLOMETRIC METHODS. Bolometer detectors containing thermal resistors which are less sensitive and require higher biasing powers than those in Table 2 (references 7, 8, 27) can be used to measure powers up to 100–200 mw. In general the methods are the same as those outlined above. Alternatively, fixed or variable r-f attenuation of known value can be interposed between the source and a low-power bolometer detector; this method is the more flexible since power levels are restricted only by the power-handling capability of the attenuator. Since attenuators capable of dissipating power up to several hundred watts (up to 1000 watts in some designs) have been realized in practice, the power range inherent in this method is at least 0–100 watts.

The attenuator's attenuation may be a significant function of frequency so that a calibration should be made (or obtained from the known frequency characteristic) at the measurement frequency. Interaction between the impedance of the source and the impedance at the attenuator's input, and between the impedance of the detector and the impedance at the attenuator's output, creates two sources of uncertainty in power measurement (see Fig. 7). These impedance interactions are preferably reduced or eliminated by an attenuator design which gives nearly unity match when alternate terminals are terminated in unity match.

Errors from attenuator calibration instability with respect to time, ambient temperature, power input level, and normal handling can usually be reduced to insignificance by (1) selection of an appropriate type of attenuator, (2) careful attention to design and construction details, and (3) frequent recalibration when necessary.

The most stable fixed coaxial attenuator types have been (1) resistance-film center conductor, (2) lumped-element π or T pad, and (3) lossy dielectric. The first two types are generally stable up to 1 watt; using lossy ceramics and heat-radiating fins type (3) can be designed to handle 100–1000 watts stably. Lossy-dielectric flexible cables have been generally unsatisfactory. In wave guides, resistance-film (parallel to the electric field) types are used up to 1 watt, and lossy ceramic dielectric types up to 100 watts. In many respects the most satisfactory fixed attenuators (in either coaxial or wave guide) have been directional couplers. Attenuation stability is unquestionably high. Attenuations above 10–15 db give the best impedance characteristics. Stringent directivity requirements are not necessary. Power-handling capability is limited solely, and input impedance primarily, by the main line termination, which may be a useful load or a dummy load of the lossy dielectric type (reference 15) or, for low powers, a resistance-film termination. The output impedance is governed primarily by the low-power termination in the auxiliary guide. It is practical to limit input and output impedance mismatches to 1.1 and 1.05 *VSWR* in cent higher than that of the inlet water and the rate of flow is m grams per second then $w = 4.18m\ \Delta t$ watts.

Temperature rise is most sensitively indicated on a microammeter deflected by d-c current generated by the differential action of series-connected thermojunctions placed alternately in the inlet and outlet water. The rate of flow is kept nearly constant during a measurement period by having a constant-head water source which is most simply a continuously refilled container of water mounted a fixed height above the water load. Calibration of the overall sensitivity can be accomplished by dissipating measurable d-c or low-frequency power in a resistor which is immersed in, and therefore heats, the water. A rate of flow different from that during measurement necessitates a proportional correction factor, so that relative flow rates must be checked. Alternatively, r-f power can be continuously compared to the low-frequency calibrating power by using a balanced bridge made up of thermojunctions in the water on each side of the calibrating resistor and on each side of the r-f water load; power measurement is then independent of water flow rate. Uncertainties about the effects of spurious heat conduction must be eliminated by having adequate thermal insulation between hot and cold junctions. Errors from heat lost by air conduction, and thermal conduction down the line, are minimized by keeping the temperature rise low (order of 1 deg cent), which requires a high rate of flow. Unfortunately this reduces sensitivity so that either a large number of thermojunctions or a sensitive microammeter is required to avoid meter-reading error. For example, a specific design using 16 hot and 16 cold junctions gave a 60-microampere deflection on a 10-ohm meter for an input power of 20 watts when the rate of flow was 3 cc/sec.

High-Q matching transformers such as the simple, single, quarter-wavelength dielectric retainer illustrated in Fig. 10, can match the water load to the transmission line satisfactorily over only a narrow frequency band (*VSWR* within about 1.2 over ±3 per cent) (reference 7), and the internally created high standing waves may cause arcing on high powers. Low-Q, wide-band, matching methods are usually preferable and may be required. One such method is to couple the main transmission line (coaxial or wave guide), by means of quarter-wavelength spaced holes (or slits) of progressively increasing sizes, to an auxiliary line containing a longitudinal dielectric tube through which the water flows; this tube forms the center conductor in coaxial line and is axially centered in wave guide. Or a tapered water termination can be simply accomplished in wave guide by mounting the wave-guide transmission line at a slight angle with respect to the horizontal.

Accuracy. Reflection loss and other r-f errors can be held to insignificant values by suitable design. Power-measurement accuracy is governed primarily by errors in calibrating standards, meter reading, and thermal loss. These errors are reduced in practice to values such that accuracies of 0.25–0.35 db are commonly attained for measurements of power between 20 and 100 watts.

COMPARISON OF CALORIMETER AND BOLOMETER-ATTENUATOR POWER READINGS. A calorimeter and bolometer-directional coupler attenuator combination can be simultaneously excited by the same source; it is thus possible to compare readings directly. Both devices, when carefully designed, are capable of such high accuracy that cross-checks usually show differences no greater than the possible inaccuracy of the attenuator calibration. If the calorimeter is the more reliable it can be used to calibrate the bolometer-attenuator combination, which can then be used with equivalent accuracy and greater convenience to measure high powers.

34. ATTENUATION MEASUREMENTS

Attenuation is defined as the ratio of the input to output power levels in a network when it is excited by a matched source and terminated in a matched load. When the latter specifications are met by the measuring circuit, uncertainties in the measured quantity are avoided, but in actual use neither the source nor the load may be exactly matched. To prevent uncertainties in its action under such conditions (see Fig. 7), accurately calibrated attenuators are usually designed to present nearly matched impedances when alternate terminals are terminated with matched loads.

Attenuation is most accurately measured by insertion methods. Readings are obtained first without the unknown, then with the unknown inserted between a source and an r-f detector (or mixer). To avoid impedance interaction errors (1) the r-f detector must be well matched (by preceding it with a well-matched attenuator if necessary), (2) the source oscillator must be adequately decoupled, and (3) the source impedance must be well matched. The last condition can be assured and additionally the source may be kept stable during the measuring period by monitoring its output with a directional coupler (see Fig. 4).

Three methods for accurate attenuation measurement will be discussed: (1) comparison against the calibrated law of a bolometer power meter, (2) comparison against a calibrated i-f attenuator in a heterodyne receiver, and (3) comparison against a calibrated or known r-f attenuator.

BOLOMETER POWER METER METHOD. The source is first terminated by the bolometer detector, then by the attenuator whose output is terminated by the bolometer detector. Attenuation is calculated from the ratio of the first to second power readings on the power meter. Since only a ratio is involved, absolute powers need not be known; therefore measurement accuracy is primarily limited by (1) meter-reading accuracy, (2) bolometer detector law calibration, and (3) temperature (zero) drift. Limitation (1) is the most important in measuring low attenuations (0–3 db) for which accuracies of ± 0.05 db are practical. The accuracy of 3–13 db measurements is limited by (1) and (2) to about ± 0.1 db. By using input and output power levels between 10 and $1/2$–$1/10$ mw, attenuations of 13–20 db are measurable within ± 0.1 to 0.2 db accuracy which is controlled by all three limitations.

HETERODYNE RECEIVER METHOD. In the heterodyne receiver method the input to the second detector is kept constant by using a calibrated attenuator preceding the intermediate frequency band-pass amplifier. I-f band widths of 0.1 to 3 megacycles with center frequencies of 30 or 60 megacycles are commonly used. Attenuation is deduced from two settings of the i-f attenuator, one with and one without the unknown between the source and the mixer. Differential frequency stability requirements can be met by incorporating an automatic frequency control circuit, or, more simply, they can be materially reduced by using a frequency-modulated mixer oscillator. Correct difference frequency occurs simultaneously with peak output which is displayed on an oscilloscope or a d-c meter preceded by a peak rectifier. Sweep methods permit the alternative use of video instead of i-f amplification and calibrated attenuation.

The minimum signal which is accurately discernible in the presence of noise interference, and the maximum signal on which the mixer is linear, usually bracket the maximum measurable attenuation to about 50–70 db; such high-level differences cannot be accurately measured unless high-level radiation and low-level pick-up are eliminated by carefully shielding all joints in the transmission line and r-f components. For example, oscillator tubes must be mounted in shielding containers into which power is supplied through r-f filters (reference 17). Maximum signal levels for which the mixer operates linearly can be determined by comparing its law against that of a bolometer power meter over a single or consecutive 10–13 db ranges. In the linear region a power meter, monitoring and indicating relative r-f signal levels, can also be used to calibrate or check the i-f attenuator. Attenuations between 0 and 60 db are measurable with an accuracy of ± 0.1–0.2 db with heterodyne receiver methods.

R-F ATTENUATOR METHOD. In the calibrated r-f attenuator method all questions about linearity are eliminated, since the input power to the first detector (or mixer) is held constant. Attenuation of the unknown is deduced from two settings of the r-f attenuator, one before and after insertion of the unknown.

A calibrated variable r-f attenuator may be a resistance-film type (reference 27) (one for wave guide is illustrated in Fig. 11), variable between 0 and 40 db by a precision mechanism which can be capable of a combined setting and reading accuracy of ± 0.1 db. Such attenuators must be calibrated and should be checked periodically by some other, more fundamental method, which may introduce an additional error of ± 0.1 db. However, the importance of this lack of accuracy is often outweighed by considerations of the convenience inherent in (1) measurement of unknown attenuation by r-f attenuator comparison methods, and (2) use of resistance-film attenuators in them.

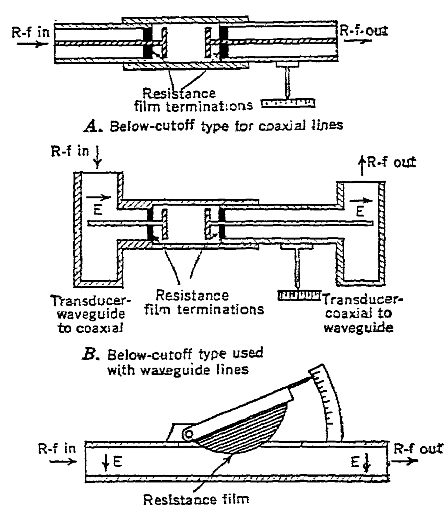


FIG. 11. Variable Attenuators

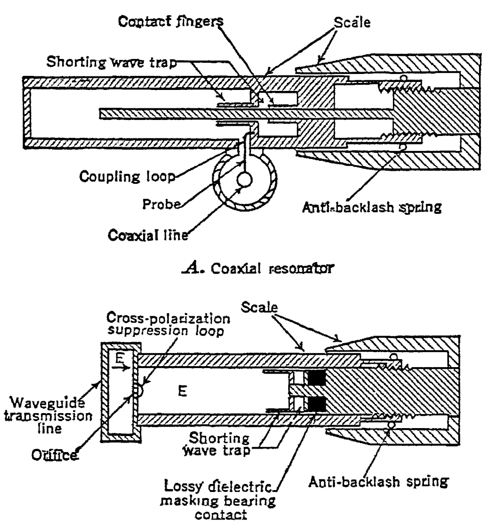
The most stable variable attenuator standards are of the wave-guide-below-cutoff type, and these are the only types for which incremental attenuation ratios can be calculated in advance (references 21 and 15). Measurement of unknown attenuation by comparison against such an attenuator is therefore a fundamental method. Unfortunately the minimum loss, above which the attenuation law is calculable without uncertainty, is around 20–30 db and may be higher if input and output impedances are corrected by using masking attenuation (reference 27). Below-cutoff attenuators therefore require the use of high-sensitivity indicators which may be (1) heterodyne receivers or (2) simply low-frequency band-pass amplifier detectors for which the signal must be amplitude modulated. The noise level may be higher in the latter system, but it is more conveniently used since only one r-f signal source is required. Accuracy is primarily limited by (1) imperfections in the attenuator mechanism, (2) indicator time stability, and (3) impedance interaction. In practice these factors have been controlled sufficiently well to allow ±0.2 db accuracy for 0–40 db attenuation measurements.

35. FREQUENCY MEASUREMENTS

Two general methods can be used to measure frequency: (1) direct comparison with harmonics of a known frequency source (reference 22) (heterodyne frequency meter method), and (2) resonance in a tunable resonator (wavemeter method). Except where the highest accuracy is needed, method (2) is the more practical since measurements can be made quickly and with the minimum of equipment. A transmission line propagating the unknown frequency is coupled to the resonator which is adjusted to resonate at that frequency; resonance is indicated by the response of a detector of relative power level in a load which terminates the transmission line.

Transmission-line resonators can be constructed to be self-calibrating by incorporating enough linear motion so that the positions of several resonances separated by half-wavelengths can be measured (lecher wire method). However, the calibration accuracy is only about 1 part in 10^3. The more accurate practice is to incorporate just enough travel to cover the desired frequency band and to calibrate the resonator by method (1) above; in this manner absolute accuracies of 1 part in 10^4 can be obtained.

RESONATORS. Resonators can be divided into two general classes: (1) resonant transmission lines (coaxial or wave guide), and (2) resonant cavities. Resonant transmission lines are shorted at one end and tuned by adjusting the line length to the other end where it is open or short-circuited. Resonant cavities are tuned by adjusting a reactance (physical discontinuity) in the electromagnetic field. On transmission-line resonators the calibrated line length changes almost linearly with transmission-line wavelength (reference 1); one specific type of cavity resonator, really a hybrid combination of TMO10 cavity and coaxial line, can be constructed to have a nearly linear calibration versus frequency over a band width as great as 12 per cent (reference 24).



B. TE11 mode resonator

FIG. 12. Wavemeter Resonators

Table 4 illustrates underlying requirements to be met for accurate frequency measurement by listing the general characteristics of two of the simplest types of transmission-line resonators (Fig. 12). These resonators are designed for approximately optimum performance when diameters are limited so that only the dominant mode (reference 1) (Section 7) can propagate. Coaxial-line resonators are generally satisfactory in the 1500–10,000 megacycle range; wave-guide resonators are superior in the 10,000–25,000 megacycle range where the Q of coaxial-line resonators become too low. In this latter range wave-guide resonators using the TE01 mode and special mode suppression techniques can be used to attain higher realizable Q's (see Section 7), but the advantage will be slight unless corresponding improvements can be introduced to reduce errors caused by imperfections in the tuning mechanism.

WAVEMETERS. The combination of a resonator coupled to a section of transmission line (to which an unknown frequency source and a detector can be coupled) will be designated as a wavemeter. Wavemeters are classified according to the way the resonator tuning affects transmission: (1) transmission type, and (2) suppression type. For the first type, the resonator must contain two orifices, probes, or loops, which allow the resonator to be connected in tandem with the transmission line; at resonance, power is transmitted through the resonator, and resonance is indicated by a maximum response on the indicator. The second type employs a single orifice, probe, or loop, to couple the resonator's impedance into the transmission line. The coupled impedance tends to suppress the resonant frequency which is indicated by a minimum response on the detector.

Table 4. Typical Wavemeter Characteristics

Frequency Band Limits	Theoretical Q Factor	Realizable Q Factor	Possible Errors from Imperfect Selectivity, Imperfect Tuning Mechanism, Impedance Mismatch			Quarter Wavelength	Inside Dimensions: Diameters and Lengths		
Mc	Q	Q_{ab}	Δfq Mc	Δfm Mc	Δfz Mc	N	d_1 in.	d_2 in.	l in.
Coaxial line									
1,500	9,200	8,000	0.01	0.03	0.01	3	0.65	2.35	5.90
2,500	11,500		0.02	0.07	0.01				3.54
3,600	7,500	7,000	0.04	0.09	0.03	5	0.32	1.20	4.14
5,000	8,400		0.05	0.2	0.04				2.95
8,500	5,800	3,500	0.2	0.2	0.1	11	0.16	0.59	3.83
10,000	6,200		0.2	0.3	0.1				3.24
19,100	4,300	1,800	0.8	0.6	0.5	19	0.08	0.29	3.10
20,000	4,400		0.9	0.7	0.6				2.80
22,800	3,800	1,200	1.5	0.8	1.0	21	0.06	0.23	2.72
25,000	3,900		1.6	1.0	1.0				2.48
TE11 mode round wave guide									
2,060	23,000	20,000	0.01	....	0.01	2	......	3.62	7.3
2,500	25,000			0.03					3.6
4,390	19,000	17,000	0.02	0.2	0.01	4	......	1.81	5.5
5,000	21,000			0.6					3.6
9,050	15,000	10,000	0.08	0.07	0.05	6	......	0.90	3.6
10,000				0.15					2.7
18,700	11,000	7,000	0.2	0.2	0.1	10	......	0.45	2.7
20,000				0.4					2.3
23,700	10,000	5,000	0.4	0.3	0.2	12	......	0.36	2.5
25,000				0.5					2.1

Transmission-type wavemeters are not favored for general frequency measurement, primarily as a result of the characteristic lack of energy transmission unless the tuning is adjusted for near-resonance. Failing to find a response on the detector, the operator is faced with uncertainty as to the adequacy of the power output of the source or the sensitivity of the detector, and this is especially troublesome if either or both must also be tuned. A suppression type is capable of just as accurate frequency measurements, and it is more convenient to use.

FREQUENCY MEASUREMENT WITH SUPPRESSION-TYPE WAVEMETERS. The simplest arrangement for measuring frequency on a transmission line is that shown in Fig. 13*A*. In the process of measuring frequency the impedance to the source is altered, and this may change its frequency unless avoided by keeping adequate attenuation between the oscillator and wavemeter. Normal operating conditions are restored by detuning the resonator. In good designs a detuned wavemeter should not create a *VSWR* in excess of 1.1. This circuit finds a wide application where only a spot check on the frequency of low-power sources is needed. Figure 13*B* shows an arrangement for continuously monitoring frequency with a negligible disturbance of energy level and impedance conditions in the main transmission line. In some cases it may be more practical to couple the auxiliary to the main transmission line by means of a single orifice, probe, or loop, but a directional coupler is preferred because less masking attenuation is needed for the wavemeter to operate out of a matched impedance, a condition for which it is calibrated.

Wavemeters should be calibrated and operated only in nearly matched circuits because reactances introduced by the source, or load, or both, change the frequency of minimum response. Using the lumped-element equivalent circuits in Fig. 14 for analysis, the magnitude of frequency change can be estimated from: $8\Delta f_z Q_{ab} = f(B_s + B_l)$, where the mismatched source and load introduce impedances having normalized susceptance components B_s and B_l at the plane of the resonator coupling. By using masking attenuators wherever necessary, it is usually practical to limit both the source and load impedance mismatch to 1.2 *VSWR*; highest value for $(B_s + B_l) = 0.4$. The errors in Table 4 were computed on this assumption and on the assumption that the wavemeter was calibrated in a perfectly matched transmission line.

The mechanical system used for tuning and indicating the resonant frequency contains potential sources of predominant error. The error in interpreting resonant frequency from a scale reading can be estimated from: $\Delta f_m = \pm M\Delta l$, where M is the rate of change of resonant frequency with tuning plunger movement in megacycles per inch and Δl is

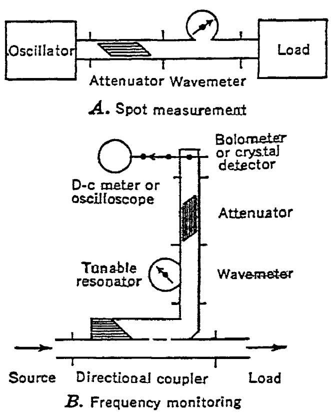


FIG. 13. Frequency Measurement Circuits

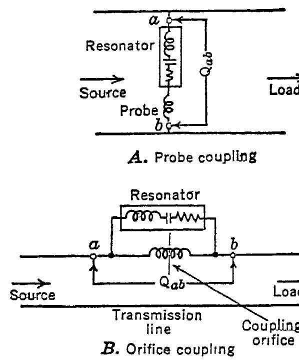


FIG. 14. Approximately Equivalent Lumped-element Circuits

the maximum possible sum of (1) backlash between the driving mechanism and the scale, (2) non-linearity in driving mechanism between calibration points, and (3) uncertainty in reading the scale. In Table 4, Δl was assumed as 0.0001 in.; even with this high order of mechanical precision the errors are large. These errors can be lowered by reducing M (band spreading). In resonant-line resonators this can be accomplished by increasing n, the number of quarter or half wavelengths. However, a compromise is usually made between tolerable error and the frequency range for which all resonant settings are unique. An alternative method for band-spreading wave-guide resonators is to operate near cut-off frequency for the resonant mode (reference 1); but this restricts band width more than the first method.

Both the resonator's Q factor and the tightness with which it is coupled to the transmission line influence the selectivity of response in the detector. Using the circuits in Fig. 14 for analysis it can be shown that the possible error in tuning the wavemeter resulting from inability to discern the absolute minimum response on a power indicator is:

$$2\Delta f_Q Q_{ab} = f\left[\frac{\Delta P}{P_m}\frac{A-1}{A^2}\right]^{1/2} \tag{3}$$

where Q_{ab} is the Q of the total impedance between terminals a–b (not the "loaded Q"); Q_{ab} is always less than the theoretical Q_1 (reference 1), and typical values for copper resonators, based upon experimental data ($A = 2$), are shown in the table. The suppression loss ratio, A, is the ratio of the power in the load when the resonator is completely detuned, P_m, to the power in the load when the resonator is tuned. The minimum discernible change in power expressed as a fraction of the detuned power is $\Delta P/P_m$. This equation is accurate only for $\Delta P/P_m$ below about 10 per cent. Assuming that P_m gives full-scale deflection and that a power change of 0.5 per cent of full scale can be discerned, Table 4 gives the possible errors for some typical resonators.

Differential Accuracy. For well-constructed wavemeters the exact calibration curve is everywhere smooth and regular. If terminating reactances remain constant the accuracy in measuring small frequency differences is primarily limited by the last two factors above.

Each one of these errors can enter twice. According to Table 4, differential measurements can be made with an accuracy no worse than about 1 part in 10^4. Accuracies of 5 in 10^5 are commonly realized in practice.

Wavemeter Calibration. Three somewhat different methods can be used to calibrate the wavemeter against low-frequency standards. In the first method the wavemeter is adjusted to resonate at a c-w frequency supplied by a local microwave oscillator; this frequency is then measured with a precision heterodyne frequency meter (accurate to 1 part in 10^6) such as shown in reference 22.

In the second method, microwave frequencies standards are harmonically generated from low-frequency standards; a calibration is then made by adjusting the frequency meter to resonate on these standards (reference 27). The response must be detected in a high-gain low-noise detector such as a double-detection receiver, since the harmonics power output is very low.

The third method is illustrated in Fig. 15, which combines advantageous features of both the above methods. The microwave oscillator is frequency modulated by a saw-tooth wave on its repeller (references 9 and 10). Its center frequency is adjusted until a low-amplitude marker oscillation appears superimposed on the rectified envelope of the oscillator output; this pip appears during the time interval for which the frequency zero-beats (2–50 kc) with a standard frequency generated by the crystal rectifier multiplier. The frequency meter is tuned until its null brackets the marker pip. Frequency-time stability of the f-m oscillator is unimportant, and the low-level output from the harmonic generator is used only for frequency marking. Additionally, a known c-w calibrating frequency can be obtained by gradually reducing the sweep to zero while keeping the marker in the center of the mode; a beat note in the head phones can be tuned in and retained by trimming adjustments on the frequency.

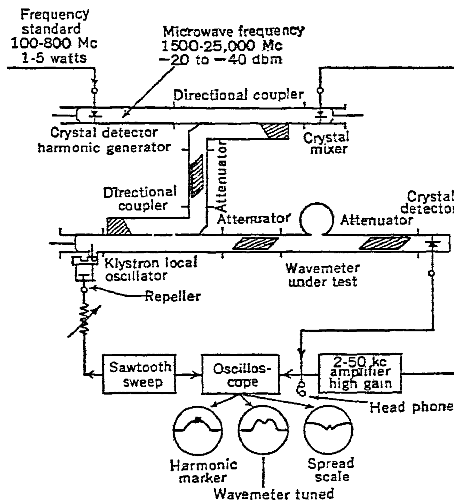


FIG. 15. Wavemeter Calibration Circuit

Conventional methods employing a master crystal-controlled oscillator and vacuum-tube multipliers can be used for generating the standard frequencies up to 800 megacycles (see p. 7-92). These frequencies are then multiplied 20 to 60 times in a silicon crystal rectifier mounted in the microwave transmission line. The harmonic number can be easily identified by a roughly calibrated or self-calibrating wavemeter. High burn-out crystals (reference 6) are recommended since they handle more input power. The microwave signal power so produced is of the order of −20 to −50 dbm. A severe requirement is therefore placed upon the noise figure of the high-gain detection system used to amplify the response to a usable amplitude.

Absolute Accuracy. By the above methods standard frequencies can be known to an accuracy of 1 part in 10^6. Adding to this the accuracy with which these standard frequencies can be transferred to the wavemeter calibration curve (limited primarily by the first two errors listed in Table 4) results in a typical absolute calibration accuracy of about 5 parts in 10^5. Then, assuming that corrections for temperature and humidity can be made with negligible error, adding all three errors in Table 4 to the calibration accuracy, a typical wavemeter can be used to measure frequency with an absolute accuracy of 1–2 parts in 10^4.

Correction for Temperature and Humidity. In order to interpret the resonant frequency from a scale reading and the calibration chart accurately, corrections must be made to allow for the effects of any change from the reference temperature and humidity conditions under which the resonator was calibrated. Two effects operate independently, and the corrections for them can be computed separately, then added algebraically: (1) temperature changes induce thermal expansion or contraction of the resonator's internal dimensions; (2) combined temperature and relative humidity changes vary the dielectric constant of the air inside unsealed resonators (reference 28).

The correction for thermal expansion of homogeneous resonators can be computed from: $\Delta f = -Cf_0\,\Delta t$, where Δf is the frequency correction to be added to the resonant frequency

f_0 indicated by the calibration chart, Δ_t is the change in temperature from that for which the calibration chart was prepared, and C is the coefficient of linear expansion for the resonator material. The correction is small for resonators made of Invar; $\Delta f/f_0$ is about $\pm 5.5 \times 10^{-7}$ for ± 1 deg fahr temperature change. For the same temperature change $\Delta f/f_0$ is about $\pm 6.6 \times 10^{-6}$ for steel resonators, and $\pm 1 \times 10^{-5}$ for brass resonators.

Corrections for the effect of dielectric constant changes are given in Table 5, which applies for unsealed resonators operated at sea-level atmosphere (reference 27). The corrections are to be algebraically added to the indicated resonant frequency (assuming that the calibration was made, or normalized, for 25 deg cent and 60 per cent relative humidity conditions). It will be observed that for ± 10 per cent change in relative humidity the correction is $\pm 1.3 \times 10^{-5}$ at 25 deg cent.

Table 5. Resonant-frequency Correction for Humidity Changes in Unsealed Resonators

Temperature, deg cent	Relative Humidity, per cent									Temperature, deg fahr
	20	30	40	50	60	70	80	90	100	
0	+0.0050	+0.0045	+0.0042	+0.0050	+0.0038	+0.0035	+0.0032	+0.0028	+0.0025	32
5	+0.0052	+0.0048	+0.0042	+0.0040	+0.0035	+0.0032	+0.0027	+0.0022	+0.0018	41
10	+0.0055	+0.0049	+0.0042	+0.0038	+0.0031	+0.0025	+0.0020	+0.0013	+0.0008	50
15	+0.0055	+0.0048	+0.0041	+0.0032	+0.0024	+0.0015	+0.0010	+0.0001	−0.0006	59
20	+0.0055	+0.0045	+0.0035	+0.0025	+0.0015	+0.0005	−0.0006	−0.0015	−0.0025	68
25	+0.0055	+0.0041	+0.0028	+0.0013	+0.0000	−0.0013	−0.0025	−0.0040	−0.0055	77
30	+0.0050	+0.0032	+0.0025	−0.0003	−0.0018	−0.0038	−0.0052	−0.0071	−0.0080	86
35	+0.0045	+0.0023	+0.0002	−0.0022	−0.0045	−0.0067	−0.0089	−0.0100	−0.0133	95
40	+0.0030	+0.0010	−0.0018	−0.0047	−0.0075	−0.0105	−0.0131	−0.0158	−0.0188	104
45	+0.0027	−0.0007	−0.0043	−0.0078	−0.0115	−0.0150	−0.0186	−0.0221	−0.0256	113
50	+0.0014	−0.0031	−0.0075	−0.0119	−0.0163	−0.0209	−0.0252	−0.0295	−0.0340	112

Frequency correction, per cent to be added to indicated reading, for resonators calibrated at 25 deg cent and 60 per cent relative humidity.

Frequency Tuning. Suppression loss ratios lower than 2 increase the difficulty in finding the region of resonance when the response is indicated on a d-c meter; since meters are inherently sluggish the narrow resonant region can be easily missed. For example, the resonant region in terms of band width between the two frequencies for which the power change is half the total null change is only $\Delta f Q_{ab} = f\sqrt{A}$. On wide-range wavemeters it is often desirable to keep this resonant region relatively wide by purposely designing for only a moderate value of Q_{ab}.

High-suppression moderate-Q requirements are relatively unimportant when a-m (including pulsed) wave envelopes are displayed on an oscilloscope screen. Since oscilloscopes respond to rapid level changes a reaction can usually be discerned, even though the resonant region be tuned through rapidly. In some cases suppression-loss ratios as low as 1.1 to 1.2 may be desirable in order to avoid excessive distortion of the oscilloscopic pattern in the region of resonance.

The oscilloscope pattern in Fig. 16 illustrates the superposition of a wavemeter null ("pip") at two frequencies on the envelope of the output of a klystron oscillator which is being frequency and amplitude modulated by a sawtooth voltage on its repeller (reference 9). The "pip" can so easily be located on this type of wave that auxiliary facilities comprising an adjustable sweep voltage and a crystal detector with oscilloscope indicator provide a convenient method for quickly tuning klystron sources to required c-w frequencies (as set on the wavemeter). The initial sweep voltage is large enough to sweep through an oscillation mode completely, regardless of the initial settings of the oscillator tuning controls; tuning is adjusted until the wavemeter pip appears at the top of the mode. The sweep voltage is then gradually reduced to zero and the oscillator tuning controls are simultaneously adjusted to keep the "pip" centered on the mode.

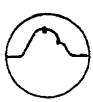


FIG. 16. Wavemeter "Pips" on Klystron Mode

BIBLIOGRAPHY

1. Schelkunoff, S. A., *Electromagnetic Waves*, D. Van Nostrand Co., New York, 1943.
2. Everitt, W. L., *Communication Engineering*, McGraw-Hill Book Co., New York, 1937.
3. Montgomery, C. G., *Techniques of Microwave Measurements*, McGraw-Hill, New York, 1947.
4. Smith, P. H., An Improved Transmission Line Calculator, *Electronics*, January 1944.
5. *Microwave Transmission Design Data*, Publication 23-80, Sperry Gyroscope Co., 1944.
6. Stephens, W. E., Crystal Rectifiers, *Electronics*, July 1946, p. 112.

7. Clayton, R. J., et al., Radio Measurements in the Decimeter and Centimeter Wavebands, *J. I.E.E.*, Part II, Vol. 93, 97 (March 1946).
8. Becker, J. A., et al., Properties and Uses of Thermistors—Thermally Sensitive Resistors, *Bell Sys. Tech. J.*, January 1947.
9. Pierce, J. R., Reflex Oscillators, *Proc. I.R.E.*, Vol. 33, 112 (February 1945).
10. Genzton, E. L., and Harrison, A. E., Reflex-klystron Oscillators, *Proc. I.R.E.*, Vol. 34 (March 1946).
11. Fiske, J. B., et al., The Magnetron as a Generator of Centimeter Waves, *Bell Sys. Tech. J.*, Vol. 25 (April 1946).
12. Hassell, F., and Jenks, F., Electroforming Microwave Components, *Electronics*, March 1946.
13. Altar, W., et al., Probe Error in Standing-wave Detectors, *Proc. I.R.E.*, Vol. 34, 33P (January 1946).
14. Mumford, W. W., Directional Couplers, *Proc. I.R.E.*, Vol. 35, 160 (February 1947).
15. Green, E. I., et al., Microwave Radar Testing, *Trans. A.I.E.E.*, Vol. 65, 274 (May 1946).
16. Whinnery, J. R., et al., Coaxial Line Discontinuities, *Proc. I.R.E.*, Vol. 32, 695 (1944).
17. Ragan, G. L., *Microwave Transmission Circuits*, McGraw-Hill, New York, 1948.
18. Tyrrell, W. A., Hybrid Circuits for Microwaves, *Proc. I.R.E.*, Vol. 35 (November 1947).
19. Type 821 Barretter, *Technical Data Bulletin*, Sperry Gyroscope Co.
20. Guillemin, E. A., *Communication Networks*, John Wiley & Sons, New York, 1949.
21. Lindner, E. G., Attenuation of Electromagnetic Fields in Pipes Smaller Than Critical Size, *Proc. I.R.E.*, Vol. 30, 554 (December 1942).
22. Essen, L., and Gordon-Smith, A. G., The Measurement of Frequencies in the Range 100 Mc/s to 10,000 Mc/s, *J. I.E.E.*, Vol. 92, Part III, 291 (December 1945).
23. Essen, L., The Design, Calibration, and Performance of Resonance Wavemeters for Frequencies between 1000 and 25,000 Mc/s., *J. I.E.E.*, Vol. 93, Part IIIA, No. 9, 1946.
24. Essen, L., Cavity-resonator Wavemeter—Simple Types of Wide Frequency Range, *Wireless Engineer*, Vol. 23, 126 (May 1946).
25. Swedin, M., Directive Couplers in Wave Guides, *J. I.E.E.*, Vol. 93, Part IIIA, No. 4, 1946.
26. Rosen, S., and Bangert, J. T., A Consideration of Directivity in Wave Guide Directional Couplers, *Proc. I.R.E.*, Vol. 37 (April 1949).
27. Gaffney, F. J., Microwave Measurements and Test Equipments, *Proc. I.R.E.*, Vol. 34, 775 (October 1946).
28. Englund, C. R., et al., Further Results of a Study of Ultra-short-wave Transmission Phenomena, *Bell Sys. Tech. J.*, Vol. 19, 369 (July 1935).

SIGNAL GENERATORS AND POWER MEASUREMENT

By F. J. Gaffney

A signal generator is a source of alternating voltage, calibrated in frequency and voltage (or power output into a specified load), and with good modulation characteristics, and carefully shielded. Units are commercially available which collectively cover frequency ranges from audio to microwave frequencies. Designs are, in general, made as broad band as the limitations of oscillator-tube and power-output calibration devices (including the output attenuator) will allow. Problems encountered in the design of signal generators are: (1) the design of stable oscillator circuits, (2) the design of systems of modulation, (3) output voltage or power standardization, (4) attenuator design, (5) shielding.

36. OSCILLATORS FOR SIGNAL GENERATOR USE

Signal generator oscillators are designed for maximum frequency stability with temperature and line voltage. Wherever possible it is desirable to load the oscillator lightly. For this reason it is advantageous at some frequencies to employ buffer amplifiers which feed the output circuits rather than to feed these circuits directly from the signal generator oscillator itself.

Tubes and Circuits. Any oscillator consists essentially of a tuned amplifier with sufficient positive feedback to supply the grid losses. The frequency stability depends on the frequency shift necessary to restore proper phase in the grid circuit when the tube characteristics change as the result of line voltage, thermal effects, etc. This frequency shift is a function of the Q of the tuned circuit employed, of the feedback system used, and of the method of coupling to the tube.

AUDIO-FREQUENCY OSCILLATORS. For the range from a few cycles to 20 kc or more, various forms of rC oscillators have proved advantageous. This type of oscillator is essentially an untuned amplifier with an rC filter providing the tuned feedback. The filter may be of the twin T or Wien bridge type. A diagram of the latter is shown in Fig. 1.

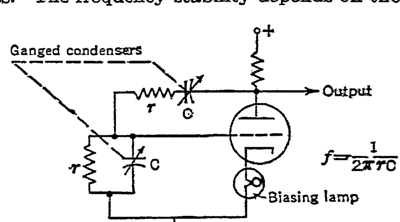


FIG. 1. Wien Bridge Oscillator

The purity of waveform depends on the sharpness of the filter, which in turn depends on the tracking of the tuning condensers. Means are usually incorporated for providing an automatic bias on the oscillator tube to insure class A operation as the oscillator output changes over the frequency band. This can take the form of a non-linear resistance such as a tungsten lamp in the cathode circuit of the oscillator. This type of oscillator if properly designed can be made extremely stable and pure in waveform. One disadvantage is the inability of a single tuning condenser to cover a large frequency range. For this reason a step switch is usually provided which switches the resistances in the circuit to provide multiple ranges.

For wider-range low-frequency oscillators (from a few cycles to several megacycles per second) beat-frequency oscillators may be used. A block diagram of such an oscillator is shown in Fig. 2. It consists essentially of two r-f oscillators, one fixed and one variable,

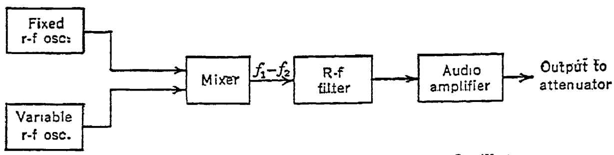


FIG. 2. Block Diagram of Beat Frequency Oscillator

which feed a mixer tube. The difference frequency between the r-f oscillators is then fed through an r-f filter to an audio amplifier. This type of circuit has the advantage that a very large frequency range can be covered with small changes in the frequency of the variable r-f oscillator. Its main disadvantage is that the audio frequency is the difference between two radio frequencies so that a small percentage variation in one of these frequencies will produce a relatively large variation in the output frequency. However, both r-f oscillators can be designed so as to be very similar in construction, and under these conditions they have a tendency to drift together so as to minimize the drift in the audio frequency. A serious problem which exists in the design of this type of oscillator is concerned with spurious outputs produced by the beating of harmonics of the r-f oscillators. These effects can be minimized by utilizing pure r-f waveforms, by the use of suitable filters between r-f and mixer stages, and by careful mixer design for large signal mixing.

RADIO-FREQUENCY OSCILLATORS. For the range of frequencies up to about 100 megacycles per second, conventional lumped constant circuits may be used. Several types of circuits such as those described in Section 7 may be employed. In order to obtain good frequency stability, either impedance-stabilized oscillators or oscillators of the electron-coupled type should be used. With the latter type of oscillator, frequency variation depends on the ratio of the screen-grid voltage to the plate voltage, and for some value of this ratio the frequency variation is extremely small with variations in plate voltage. The electron-coupled oscillator, though good from the standpoint of stability, produces a poor waveform which contains many harmonics. Care must therefore be taken in the use of such an oscillator to insure that the harmonic content does not affect the power-measuring circuit or the receiver being tested. The frequency stability of all types of oscillators is improved by regulation of the B voltage. Some frequency instability can be produced by variation in cathode heater voltage, but this is seldom compensated for in practical design. Frequency stability with variation in ambient temperature is accomplished through careful design of the tuned circuit of the oscillator. Coil forms should be wound on a material having a very low coefficient of expansion, such as Vicor. Special attention must be paid to the design of the variable condenser to minimize dimensional changes with temperature.

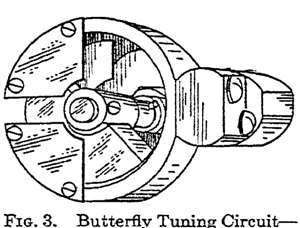


FIG. 3. Butterfly Tuning Circuit—400 to 1200 megacycles per second (Courtesy of General Radio Co.)

At frequencies above 100 megacycles per second considerable difficulty is encountered in the use of conventional lumped circuits. The losses in such circuits become excessive at these frequencies, and the tuning range becomes small, since, as the inductance is decreased to obtain higher frequencies, the tube and wiring capacitances become a larger proportion of the total allowable capacitance. One solution to this problem has been in the use of circuits of the butterfly type such as that shown in Fig. 3. Here the inductance and capacitance are varied simultaneously. This has the effect of increasing the tuning range and simultaneously maintaining the L/C ratio. Such circuits have been built for frequencies up to 2000 megacycles per second. At frequencies above 1000 megacycles per second, however, the losses are too great for practical oscillator applications.

Another approach to the problem of obtaining tuned circuits at high frequencies consists in the use of resonant sections of transmission line or in the use of cavity-type resonators. Here, radiation losses are eliminated and the distributed nature of the circuit allows currents to flow over larger areas, thus permitting low copper losses. For a lossless transmission line length l, short-circuited at the far end, the input impedance z_{in} is given by:

$$z_{in} = z_0 \tan \frac{2\pi l}{\lambda} \tag{1}$$

where z_0 = characteristic impedance of the line and λ = wavelength.

Thus, if $l = \lambda/4$, $z = \infty$. Actually, owing to losses, transmission lines dissipate energy and have a Q factor which is defined in terms of band width as in a lumped constant circuit. Transmission line Q is approximately given by:

$$Q = \frac{2\pi f_0 z_0}{rc} \tag{2}$$

where f_0 = frequency of resonance (cycles per second), z_0 = characteristic impedance = $\sqrt{L/C}$ (ohms), r = resistance per unit length (ohms/meter), and c = velocity of light (meters per second).

The input impedance of a quarter wavelength short-circuited line is then approximately given by:

$$z_{in} = z_0 Q$$

In the equation for Q, both z_0 and r vary with the dimensions of the line. For maximum Q there exists an optimum diameter ratio for coaxial lines (for conductors of the same resistivity) and an optimum ratio of spacing to the diameter of the conductors for parallel wire lines. Letting this ratio be b/a for the two cases, one value of b/a gives maximum Q and a second value gives maximum input impedance. The values, together with the corresponding characteristic impedances are given in the table.

	Coaxial Lines		Parallel Wire Lines	
	b/a	z_0	b/a	z_0
Max. Q........	3.6	76.8	4.0	851.
Max. z........	9.2	133.1	8.0	1934.

With b/a constant, both Q and z increase linearly with b. Such a line can then be used as an antiresonant circuit, with the resonance frequency determined by the line length. Actually the shapes of the curves of reactance and resistance as a function of frequency are not identical to those of lumped constant circuits but are quite similar near the resonant frequency. Parallel wire transmission lines, because of their simplicity of construction, are sometimes used in experimental oscillators, but coaxial transmission lines are more commonly used in signal generator applications because of their lower losses (the radiation loss being zero) and self-shielding construction. The only difficulty with such resonant lines is in the physical lengths required and in the mechanical difficulties with sliding contacts. For a frequency of 100 megacycles per second, for instance, a quarter-wavelength transmission line would have a length of 75 cm. The line may, however, be artificially shortened by the use of a fixed condenser across its input terminals.

Vacuum tubes designed for low-frequency applications fail at higher frequencies because of losses, limitations due to input and output capacitances, and transit time effects. To minimize these defects, special tubes have been designed for the higher-frequency ranges. By making the tube elements very small, spacings between elements can also be made small without the introduction of excessive interelectrode capacitances. Lead inductances may be minimized by bringing out more than one lead from each electrode as in some types of high-frequency acorn tubes.

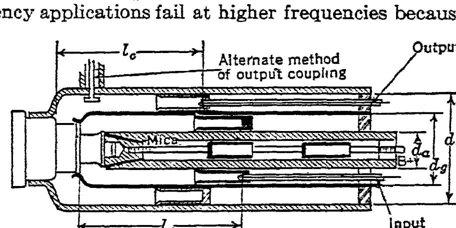


FIG. 4. Typical Double-concentric-cavity Circuit (Courtesy of General Electric Co.)

A better scheme is that exemplified by tubes of the 2C40 type which utilize disk seal construction to reduce lead inductances to a minimum. This type of oscillator tube lends itself particularly well to incorporation into a coaxial-line oscillator circuit. One such design showing a tuned plate-tuned grid oscillator is shown in Fig. 4. Owing to the difference in end effects and to the capacitive loading contributed by the tube elements, it is usually necessary to move the tuning plungers at a different rate. This causes some difficulty with mechanical tracking of the two cavities.

An alternative design known as a reentrant oscillator is shown in Fig. 5. This type of oscillator may be controlled by means of a single tuning plunger but will operate satisfactorily only over a relatively narrow frequency band. Oscillators using this type of circuit with a 2C40 triode tube have been made for frequencies as high as 3000 megacycles per second. Experimental tubes have been made to oscillate in this type of circuit at even higher frequencies. Because of the close grid cathode spacing a considerable frequency variation may be produced by variations of the tube heater voltage.

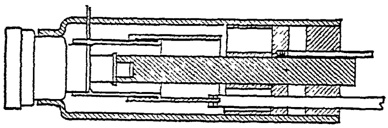


FIG. 5. Method for Tuning Re-entrant Oscillator Circuit (Courtesy of General Electric Co.)

At frequencies above 2000 megacycles per second, velocity variation tubes become useful. In this type of oscillator tube transit time is utilized to provide bunching of the electrons in a drift space. The efficiency of this type of oscillator is low and the frequency stability is relatively poor. No upper frequency limit exists for oscillators of this type except that imposed by problems of physical construction. This limitation occurs somewhere in the region of 60,000 megacycles per second. Cavity resonators which may either be external to the tube or integral with it are utilized with this type of oscillator. The type of cavity resonator used is one which develops maximum voltage across the bunching grids in the tube.

The type of velocity variation tube known as a reflex oscillator is most convenient for signal generator applications. An outline diagram of this type is shown in Fig. 6. In this tube the electron stream is velocity modulated by the bunching grids. The electrons then drift in a space between the bunching grids and a negatively charged reflector which turns them around and causes bunches to arrive again at the bunching grids in such phase as to deliver energy to the cavity resonator of which the bunching grids form a part. The resonant frequency of such an oscillator may be changed by varying the cavity dimensions. Alternatively, the frequency may be changed by changing the capacitance between the buncher grids. The latter method requires that the tube be built with a flexible diaphragm since the grids are of course in an evacuated space. The realizable tuning range with this type of tuning is only about 20 per cent. This design does, however, possess the advantage that the tube and the circuit are self-contained and troubles with sliding contacts are eliminated.

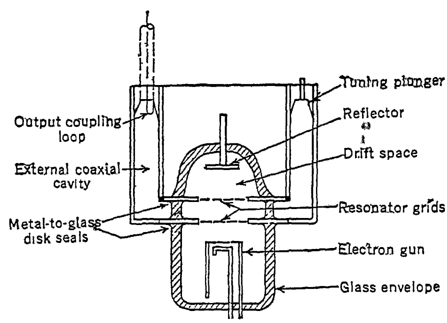


FIG. 6. Reflex Type Velocity Variation Oscillator

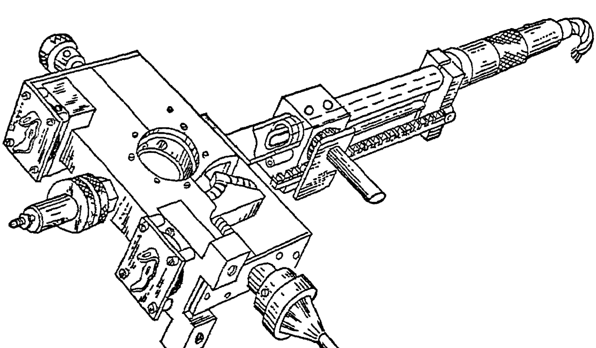


FIG. 7. TM-mode Cavity for Reflex Velocity Variation Oscillator

By bringing the bunching grids out through disk seals in a glass envelope, an external cavity may be used with this type of tube. Cavities of both coaxial mode and the TM mode have been employed successfully. A cavity of the coaxial-mode type is shown in Fig. 6; Fig. 7 shows an outline drawing of the TM-mode cavity. With the coaxial-type cavity frequency ranges of better than 2 to 1 may be achieved. Care is again necessary in the construction of tuning plungers so as to make good contact over the tuning range. In order to minimize rubbing contact, it is possible to design choke-type contacts which are themselves resonant sections of transmission lines.

37. MODULATION OF SIGNAL GENERATORS

In order to test the detection characteristics of receivers and to supply a signal that can be amplified by audio amplifier methods, signal generators are usually equipped with means for modulating the r-f voltage output. Several types of modulation are used, depending on the types of equipment with which the signal generator is most likely to be employed. Signal generators for use in the a-m broadcast band, for instance, are provided with amplitude modulation, and those for use in testing f-m receivers are frequency modulated. It is generally desirable to limit the modulation to one type and to provide means of reducing unwanted types of modulation.

AMPLITUDE MODULATION. Both the impedance-stabilized oscillator and the electron-coupled oscillator are susceptible to plate modulation with little accompanying frequency modulation. In the impedance-stabilized oscillator, the frequency is independent of applied plate voltage over a wide range if the proper value of stabilizing impedance is used. Since signal generator oscillators are tuned over wide frequency ranges, however, it is necessary to track the stabilizing impedance with the main frequency control if satisfactory performance is to be obtained. In the Hartley oscillator, the stabilizing impedance is a condenser for both the grid and plate stabilization types, and this condenser must be kept proportional to the total value of tuning condenser as the latter is varied. Figure 8 is a diagram showing this type of oscillator with plate modulation.

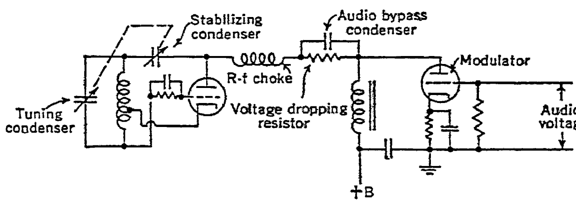


FIG. 8. Modulation of Impedance Stabilized Oscillator

The audio power output from the modulator tube must be about 3 times the r-f power from the oscillator if 100 per cent modulation is to be obtained. This can be accomplished by means of a voltage dropping resistor between the plates of the r-f and modulator tubes having a value such that the plate potential of the r-f tube is about 70 per cent that of the modulator. This resistor must be adequately bypassed for the lowest audio frequency employed. The time constant of the self-biasing circuit in the grid of the r-f oscillator must be such as to be able to follow the highest modulation frequency if the same peak audio voltage is to provide the same percentage modulation of the carrier.

The electron-coupled oscillator circuit depends for its frequency stability on the maintaining of a fixed ratio for plate and screen voltages. This is best accomplished by feeding the screen from a voltage divider between plate and ground of sufficiently low resistance that the screen voltage is invariant with screen current. The screen must be bypassed with a condenser having low impedance to the r-f voltage but high impedance to the modulating voltage. As in the impedance-stabilized oscillator, a dropping resistance must be employed between the plates of the r-f and modulator tubes if 100 per cent modulation is to be obtained. A possible variation, of course, is to supply the tubes from different taps on a power supply and to couple the modulator to the r-f oscillator by means of a transformer. A circuit diagram illustrating the former scheme is shown in Fig. 9, and the latter alternative is shown in Fig. 10. This eliminates the need of a dropping

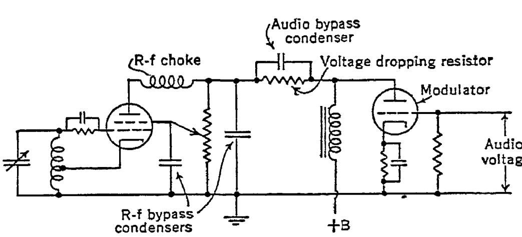


FIG. 9. Direct Coupled Modulation of Electron Coupled Oscillator

resistor but requires a tapped power supply voltage and a transformer of flat response over the audio band width used. Where extreme freedom from f-m effects is desired, an r-f amplifier is modulated rather than the r-f oscillator. Under these conditions, the r-f oscillator works at constant potential and is lightly loaded. This arrangement also makes the oscillator frequency stable with changes in loading of the output attenuator. The only disadvantage of the scheme is concerned with the necessity for providing a tuned amplifier which is ganged with the r-f oscillator. The increase in stability obtained, however, is sufficiently great to warrant the use of this method with signal generators of the precision type.

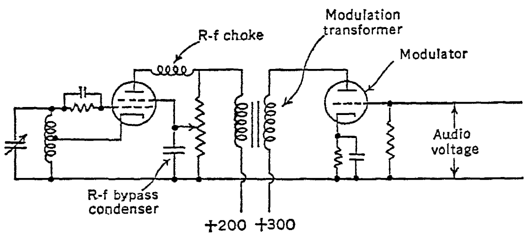


FIG. 10. Transformer Coupled Modulation of Electron Coupled Oscillator

Most signal generators for use in testing broadcast receivers are equipped with means of modulating at 400 cycles, this frequency having been standardized for receiver testing. Usually, provision is also made for modulation by means of an external oscillator operating at any desired frequency in the audio range. If suitable precautions have been taken in choosing the time constants of the grid leak and condenser biasing system and of the plate dropping networks, per cent modulation may be calculated by impressing various d-c voltages on the plate of the r-f oscillator and measuring the variation in r-f output by means of a vacuum-tube voltmeter or other method. A meter which measures the impressed audio voltage may then be calibrated in terms of per cent modulation. Various types of a-c meters have been used for this purpose such as thermocouple, vacuum-tube voltmeter, and rectifier-type meters. If other than sine wave modulation is externally applied, the indication of such meters must be corrected accordingly.

FREQUENCY MODULATION. For use in visual alignment of wide-band filters as well as for testing frequency-modulation receivers, a frequency-modulation generator is often required. Several methods for varying the radio frequency at an audio rate have been employed. The simplest of these consists of a rotating trimmer condenser, in parallel with the oscillator tank, mounted on the shaft of a small motor. This scheme allows wide frequency variation. It suffers, however, from several defects, among them being troubles encountered from vibration, contact troubles if slip rings are used for connection, variation in amount of frequency swing as the center frequency is varied, and production of undesired amplitude modulation. The audio waveform and frequency of such a device are usually fixed. A similar scheme makes use of a vibrating rather than a rotating plate. This allows modulation at higher audio frequencies and the frequency may be more readily varied. The obtainable frequency sweep is much smaller, however, and a frequency multiplier scheme must usually be employed to obtain the required sweep in the r-f output frequency.

A method which is more complex but considerably more versatile makes use of a reactance tube modulator. The frequency modulation obtainable with this method without large accompanying amplitude modulation is small, and a frequency multiplier scheme must be employed. To eliminate the necessity of tuning the multiplier stages, a heterodyne system such as that shown in Fig. 11 is usually employed.

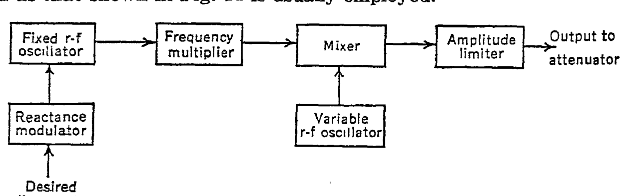


FIG. 11. Constant Deviation Frequency Modulator

With such a system, the frequency swing of the output frequency is independent of the center frequency. A typical generator of this type produces a frequency swing of over a megacycle per second at center frequencies from 60 to 120 megacycles per second.

At microwave frequencies, where velocity modulation tubes are usually employed, frequency modulation can readily be accomplished by applying the modulation signal to the reflector of a reflex-type oscillator. At 3000 megacycles per second, for example, a frequency swing of approximately 30 megacycles per second can be obtained with accompanying amplitude modulation of approximately 50 per cent. Frequency swings of 4 or 5 megacycles per second can be obtained with negligible amplitude modulation.

PULSE MODULATION. For testing radar, blind landing, and similar systems, pulse-modulated generators are required.

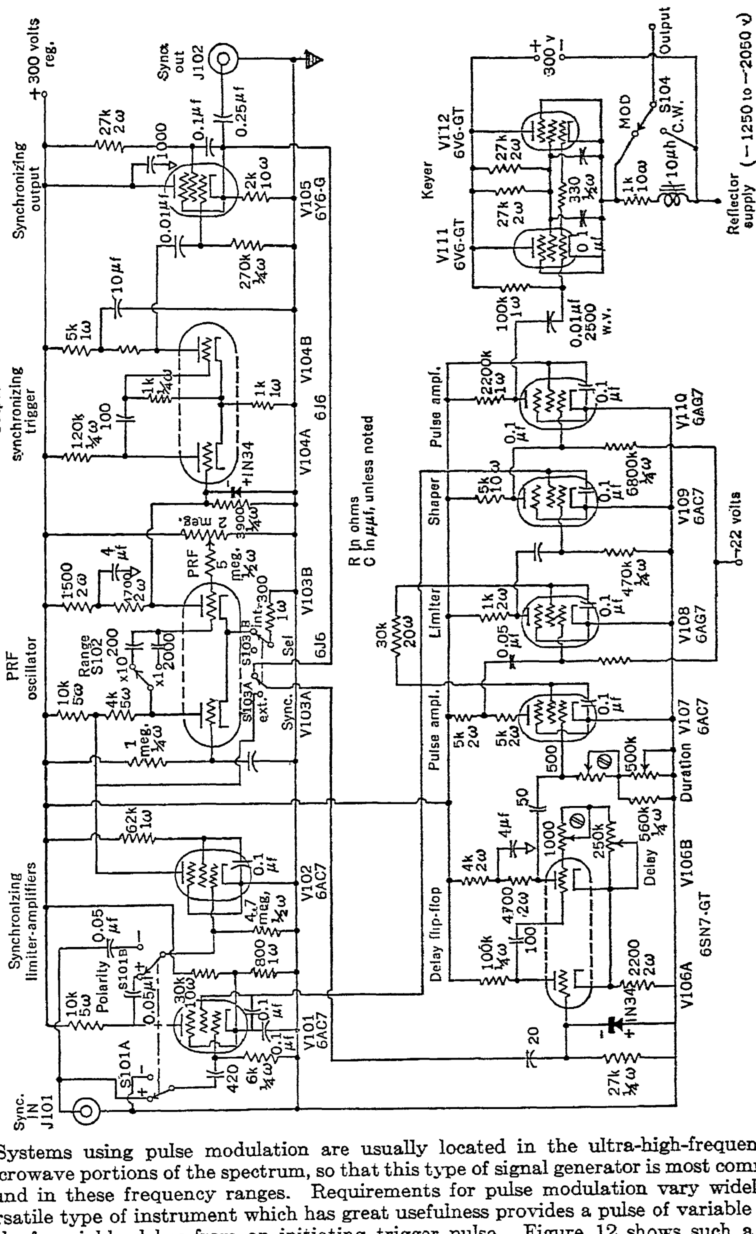


FIG. 12. Pulse Modulator

Systems using pulse modulation are usually located in the ultra-high-frequency or microwave portions of the spectrum, so that this type of signal generator is most commonly found in these frequency ranges. Requirements for pulse modulation vary widely. A versatile type of instrument which has great usefulness provides a pulse of variable width and of variable delay from an initiating trigger pulse. Figure 12 shows such a pulse modulator designed to operate with a velocity variation oscillator. The circuit is designed to operate either with an external trigger voltage applied at the *SYNC IN* jack (J-101) or with its own synchronizing signal developed by the *PRF* oscillator. In the latter case the forward edge of the square pulse produced by the *PRF* multivibrator V-103 is differentiated in the grid circuit of the output synchronizing trigger amplifier V-104 so as to produce a short pulse in the output of this tube. This pulse is then fed to a cathode follower output stage which delivers a positive synchronizing trigger from J-102. The output of either the external synchronizing pulse or the self-generated synchronizing pulse is selected by switch S103-A and fed to the delay multivibrator V-106. The trailing edge of the pulse produced in this multivibrator is differentiated in the grid circuit of the pulse amplifier tube V-107. The negative pulse so produced drives V-107 far below cutoff, and the time required to reach the conducting state again is determined by the time constant of the differentiating circuit in the grid of the pulse amplifier stage. This pulse is then squared and amplified in V-108, V-109, and V-110, and delivered to the keyer V-111 and V-112. The cathode return of the keyer tube is connected to the normal reflector supply voltage for the velocity variation tube, which may be adjusted to the correct value for proper operation of the tube when the *MOD-CW* switch is in the *CW* position. With the switch then thrown to the *MOD* position, and in the absence of a pulse, the reflector potential is such as to preclude oscillation. During the pulse, however, the reflector voltage is restored momentarily to its previous value. Pulsed operation of the velocity variation tube is thus achieved. The circuit is designed to operate with pulse recurrence rates from 50 to 5000 pps with pulses delayed from the trigger pulse from 1 microsecond to 100 microseconds, and having widths variable from 0.5 to 30 microseconds.

In using such pulse modulation systems, care must be exercised in assuring that the Q of the r-f circuit being modulated is sufficiently low to pass the sidebands generated by the pulse operation. For r-f frequencies below 100 megacycles per second this requirement is difficult to meet if stable oscillator performance is to be obtained. A method which is reasonably successful consists in pulsing an r-f amplifier driven by the oscillator. The main problem encountered here is to provide sufficient shielding so that no r-f output is present between pulses. The degree of shielding required is very great if it is desired to test sensitive receivers, and it is difficult to obtain largely because of coupling due to internal electrode capacitances of the amplifier tube. A second method consists of pulse-modulating a higher-frequency oscillator and obtaining the desired frequency by heterodyne methods. The problem in applying this technique is largely that of filtering out the undesired frequency components in the output.

38. STANDARDIZATION OF OUTPUT POWER

One of the most difficult problems in signal generator design is to determine the r-f power output accurately. One method that has seen considerable application in the lower-frequency ranges is to measure the current into a resistive output attenuator by means of a thermocouple. The quantity of interest is, of course, the output voltage (or power) into a specified load impedance. It is desired, then, to have the thermocouple read the voltage (or power) at the input to the attenuator so that, the attenuator having been calibrated, the output voltage (or power) will be known accurately for all settings of the attenuator. Several considerations enter into the validity of such a calibration.

In order that a given thermocouple reading will correspond to a given voltage or power level at the input to the attenuator, the impedance seen looking into the attenuator from the thermocouple must remain constant for all attenuator settings, assuming that the attenuator output is correctly terminated. At frequencies low enough so that the attenuator elements can be considered pure resistances, this condition can be met (see article 39). Difficulty is encountered at the higher frequencies, however, in that the attenuator input becomes reactive so that its impedance varies with frequency. Under these conditions the input current no longer bears a fixed relationship to the output voltage. At higher frequencies, the thermocouple impedance itself is also subject to variation due to the impedance presented to the heater by the couple and its associated measuring leads. The design of an output filter for the couple, which permits a very low capacitive admittance to ground at the higher frequencies, becomes very difficult.

The difficulty is minimized by using heater wires of very small diameter. Thermocouples of the separate heater type, where the couple is insulated from the heater by a small glass bead, minimize these effects. They are, in turn, sluggish in operation and less sensitive than direct-contact types. Correction must also be made for skin effect of thermocouples used at high frequencies.

At very high frequencies, where the physical distance between the thermocouple and the attenuator and between the thermocouple and the signal generator oscillator becomes an appreciable part of a wavelength, further trouble is encountered from reflections from the couple itself as seen by the signal generator oscillator and variations in the phase of such reflections as a function of frequency due to changes in the electrical lengths involved as the frequency changes. For these reasons, thermocouples have seen their greatest application in the frequency range below 10 megacycles per second although they may be used with special precautions at very much higher frequencies.

A second method for standardizing output consists in the use of a high-impedance vacuum-tube voltmeter at the input to the attenuator. It is usually desirable to have the reading of such a vacuum-tube voltmeter independent of modulation. This may be accomplished by utilizing an averaging type voltmeter circuit. This type of operation is obtained when the vacuum tube of the vacuum-tube voltmeter is operated with large voltage input and with grid leak and condenser time constant sufficiently small to allow it to follow the audio modulation. Under these conditions, the meter reads the average r-f voltage applied to its terminals. Diode voltmeters with large applied voltage may also be used; they have the advantage of maintaining calibration with tube life. It should be remembered that the power output from a modulated oscillator is proportional to $(1 + m^2/2)$, where m is the fractional modulation (per cent modulation/100), so that m must be known if it is desired to determine the power delivered to the attenuator (or effective voltage at the attenuator input).

Vacuum-tube voltmeters may be used at frequencies up to those where transit time effects reduce the rectification efficiency and thus introduce error. Transit time may be minimized by close electrode spacing and high applied voltages. Close spacing, however, introduces large interelectrode capacitance, and this may itself become a source of error. Interelectrode capacitance may be reduced by making the physical size of the electrodes small, so that an ideal tube for the purpose would have very small electrodes spaced close together. In addition, it is necessary to take special precautions to minimize the inductance of the leads to the tube electrodes. In the acorn-tube types lead inductances are sometimes minimized by bringing out two leads in parallel from each electrode. In tubes such as the 2C40, heavy cylindrical leads are employed. For acorn-type diodes, the reduction in rectification efficiency is of the order of 30 per cent at 500 megacycles per second for voltages of 0.5 volt or less. This effect, together with the change of input impedance with frequency, limits the range of usefulness of the vacuum-tube voltmeter as a power output monitor.

The deleterious effects of transit time can be minimized to some extent by substituting a crystal rectifier for a vacuum tube in a voltmeter circuit. A semiconductor such as silicon in crystalline form is embedded in a conducting material which forms one contact. The second contact is made through a tungsten whisker. The resistance between the whisker contact and the semiconductor is non-linear and may, therefore, be used to rectify an impressed alternating voltage. Though transit time in such a crystal rectifier is entirely negligible, a similar effect occurs which is due to the capacitance between the whisker contact and the crystal which may be thought of as shunting the non-linear resistance as shown in Fig. 13. Here, r_s represents the resistance of the bulk crystal between the barrier layer and the fusible metal in which the crystal is embedded, r_e represents the non-linear resistance between the whisker and the crystal, and C is the capacitance between the whisker and the crystal across the barrier layer. At high frequencies, the capacitance shunts the non-linear resistance in such a way as to reduce the rectification efficiency, producing an effect similar to that produced by transit time in tube rectifiers. With the type 1N21-B rectifier, however, this effect is small for frequencies at least as high as 1000 megacycles per second. Other disadvantages to the use of crystals exist, however, among them non-uniformity of characteristic and variation of impedance among crystals of the same type, and a change of characteristics with overload.

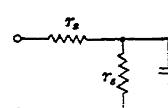


FIG. 13. Equivalent Circuit of Crystal Rectifier

For very high-frequency applications, the variation of indicated power with frequency due to variation in electrical length of the leads to the power indicator has led to the use of systems which divert a known fraction of the power from the signal generator oscillator to the power indicator. Bolometer elements are frequently used as the power-detecting devices in such a system (see article 11–33). They are sensitive detectors of r-f power, having resistance slopes of several thousand ohms per watt. For example, a 10-milliampere Littlefuse has a slope of 2800 ohms per watt, while a 5-milliampere Littlefuse has a slope of 5100 ohms per watt. Commercially available Littlefuses are often used as r-f power detectors and are quite satisfactory up to frequencies of several hundred megacycles per second. At higher frequencies, their geometry is undesirable and special bolometer elements such as that shown in Fig. 14 are employed.

The resistance-power curve of a thermistor is not accurately linear as is that of a hot-wire bolometer. This is of little importance when it is desired to monitor the r-f power at constant level but becomes important when it is desired to provide an easily calibrated indicator which will measure power over some range. The r-f impedance of a bolometer depends on its geometry as well as on its resistance. It has been found more difficult in manufacture to hold the r-f impedance of thermistor units than it is with special hot-wire bolometers.

If it is desired to operate any type of temperature-sensitive power indicator over wide ranges of ambient temperature, some form of temperature-compensating circuit must be employed. This problem becomes more acute as the sensitivity of the power indicator is increased. Temperature-sensitive elements such as those used in the power-measuring

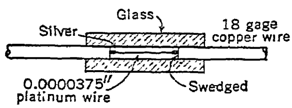


FIG. 14. Hot-wire Bolometer

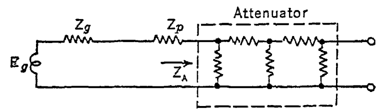


FIG. 15. Power Monitor Equivalent Circuit

circuit may be employed to accomplish the required compensation. For a direct-reading bridge circuit, two such compensating elements are, in general, used. One of these maintains the zero reading, and the other adjusts the slope of the indicator so as to maintain the correct calibration.

In the use of power indicators which absorb a fraction of the generator power, care must be taken to insure that the power split between the power indicator and the output circuit is maintained constant over the frequency band of operation.

At the lower radio frequencies where line lengths may be neglected this problem does not present serious difficulties. Here, an arrangement such as that shown in Fig. 15 may be used where the power-sensitive element is in series with the output attenuator. Here, if the line length between the power detector z_p and the output attenuator z_a can be neglected, the current in the two elements is common so that the power ratio is simply

$$\frac{P_p}{P_a} = \frac{r_e(z_p)}{r_e(z_a)} \tag{3}$$

where $r_e(z_p)$ represents the real part of the power indicator impedance.

At frequencies high enough so that the line length becomes an appreciable part of a wavelength, the current is no longer the same in the two elements and the equation holds only if the impedance of the attenuator is transformed down the line to a point close to z_p. Then, if the impedance of the attenuator is frequency dependent, changes in both the real and reactive components will affect the power split. The difficulty is overcome only by maintaining the attenuator impedance close to the characteristic impedance of the line over the frequency range.

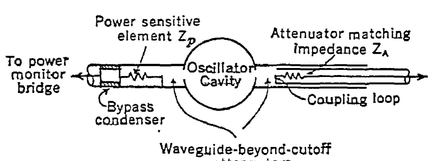


FIG. 16. Power Monitor for Cavity Oscillator

At very high frequencies, coaxial or wave-guide structures are used. One method of monitoring power from a cavity-type oscillator is shown in Fig. 16. Here the power from the oscillator is coupled through two openings in the cavity to wave-guide-beyond-cutoff attenuators which feed the power monitoring and output circuits respectively.

The power delivered to the monitor is then:

$$P_p = \left|\frac{E_p}{z_p}\right|^2 r_e(z_p) \tag{4}$$

where E_p = voltage induced in coupling loop feeding the power detector, z_p = impedance of power detector, and $r_e(z_p)$ = real part of z_p.

Similarly, the power delivered to the attenuator is given by:

$$P_A = \left|\frac{E_A}{z_A}\right|^2 r_e(z_A) \tag{5}$$

where E_A = voltage induced in coupling loop feeding the attenuator, z_A = impedance of attenuator seen looking from the loop, and $r_e(z_A)$ = real part of z_A.

The ratio of these two powers is then given by

$$\left|\frac{E_p}{E_A}\right|^2 \times \left|\frac{z_A}{z_p}\right|^2 \times \frac{r_e(z_p)}{r_e(z_A)} = \frac{P_p}{P_A} \tag{6}$$

Since the ratio E_p/E_A can be made quite constant over a wide frequency band, if symmetrical coupling to the cavity is employed, this reduces to:

$$\frac{P_p}{P_A} = K\left|\frac{z_A}{z_p}\right|^2 \times \frac{r_e(z_p)}{r_e(z_A)} \tag{7}$$

One way to make the right-hand side of this equation constant is to have both z_p and z_A equal the real characteristic impedance of the lines in which they are placed.

A similar situation obtains in wave-guide circuits where the monitor and attenuator are placed in two arms of a wavelength T and are fed from the third arm as shown in Fig. 17. Here the monitor and attenuator are effectively in series. The power split is given by

$$\frac{P_p}{P_A} = \frac{r_e(z_p)}{r_e(z_A)} \tag{8}$$

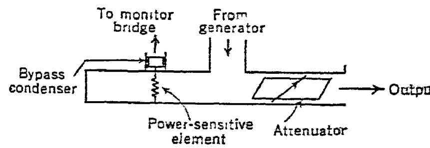


FIG. 17. Wave-guide Power Monitor

This can also be made constant by matching both attenuator and power indicator to the wave guide.

39. ATTENUATOR DESIGN

For frequencies up to about 2 megacycles per second, resistive attenuators using Ayrton Perry non-inductive wire-wound resistors may be used. In order to maintain the input and output impedance of such an attenuator constant as the attenuation is varied, iterative networks of the T or π type are used. Such a T network, designed to match without reflection the generator impedance r_g to a load resistance r_L, is shown in Fig. 18. If the generator and load are to be connected to the network without introducing reflection, r_{in} must equal r_g and r_{out} must equal r_L. The attenuation constant of such a network is defined by

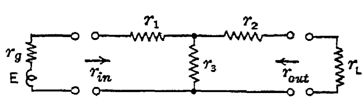


FIG. 18. T-type Attenuator

$$\frac{E_{\text{out}}}{E_{\text{in}}} = \frac{I_{\text{out}}}{I_{\text{in}}} = \varepsilon^{-\alpha} \tag{9}$$

or

$$\alpha = \ln\frac{E_{\text{out}}}{E_{\text{in}}} = \ln\frac{I_{\text{out}}}{I_{\text{in}}} \tag{10}$$

For given generator and load resistances and specified attenuation constant the elements of the network are given by:

$$r_1 = \frac{(r_g + r_L)\tanh\left(\frac{\alpha}{2}\right) + r_g - r_L}{2} \tag{11a}$$

$$r_2 = \frac{(r_g + r_L)\tanh\left(\frac{\alpha}{2}\right) - r_g + r_L}{2} \tag{11b}$$

$$r_3 = \frac{r_g + r_L}{2\sinh\alpha} \tag{11c}$$

In the special case where $r_g = r_L = r_0$, these expressions reduce to:

$$r_1 = r_2 = r_0\tanh\frac{\alpha}{2} \tag{12a}$$

$$r_3 = \frac{r_0}{\sinh\alpha} \tag{12b}$$

The unsymmetrical case is of interest in signal generators for the broadcast band since it is sometimes desired to make the output impedance sufficiently low that the voltage developed across it is independent of the impedance connected to the output terminals. The attenuation of the network may be varied and constant input and output impedances maintained if the resistances are kept in the ratio defined by the equations.

In practice, accuracy requirements and difficulty in shielding output from input make it undesirable to utilize a single T or π section for covering large ranges of attenuation.

For these reasons, a ladder-type attenuator is commonly employed. A useful form of such an attenuator is shown in Fig. 19. If K = the ratio by which the voltage across r_L is reduced in moving from one switch point to the next, the resistances r_1 and r_2 are given by:

$$r_1 = r_L\,(K - 1) \tag{13a}$$

$$r_2 = r_L\left(\frac{K}{K-1}\right) \tag{13b}$$

The input resistance r_{in} is independent of the position of the tap switch and is given by:

$$r_{\text{in}} = \frac{r_s r_L}{r_s + r_L} \tag{14}$$

where

$$r_s = \tfrac{1}{2}\,[\sqrt{r_1(r_1 + 4r_2)} + r_1]$$

Between r_L and the step attenuator, it is usually desirable to place a variable T pad attenuator of the type previously described in order to obtain a fine control. r_L may be the external load resistance into which the generator is designed to feed, or it may be made

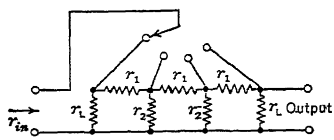


FIG. 19. Ladder-type Attenuator with Constant Input Impedance

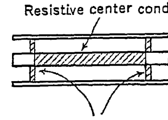


FIG. 20. Coaxial π Section

low in value and incorporated into the signal generator in which case the actual load impedance must be large compared to r_L for proper operation. For satisfactory operation, each section of the attenuator must be separately shielded and a separately shielded low-capacity switch must be employed.

For high-frequency applications, T or π sections may be built in coaxial line form as shown in Fig. 20. Units of this type may be used up to the frequency where the length of line becomes an appreciable part of a wavelength, and they have seen successful operation up to frequencies as high as 1000 megacycles per second. The resistive elements may conveniently be made of glass rods and disks coated with thin metallic films. A decade attenuator may be made by mounting several such elements in a turret arrangement.

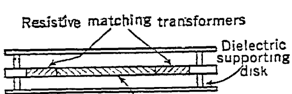
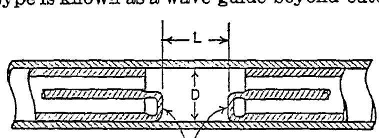


FIG. 21. Microwave Attenuator

At microwave frequencies, attenuators consisting of resistive sections of coaxial line equipped with resistive matching transformers have been successfully employed. A diagram of such a unit is shown in Fig. 21.

The band width of such devices is limited by the change in attenuation with frequency and by the frequency sensitivity of the resistive matching transformers.

A type of attenuator which has seen increasing use in signal generator applications is shown in Fig. 22. This type is known as a wave-guide-beyond-cutoff attenuator and makes



$$\text{Loop coupling attenuation} = 32.0\sqrt{1 - \left(\frac{1.71D}{2}\right)^2}\ \text{db per diameter displacement of loops.}$$

$$\text{Disk coupling attenuation} = 41.8\sqrt{1 - \left(\frac{1.30D}{2}\right)^2}\ \text{db per diameter displacement of loops.}$$

FIG. 22. Waveguide-beyond-cutoff Attenuator

use of the fact that, for diameters below a critical value which depends on the frequency and mode of propagation, waves in a hollow tube suffer no phase displacement but are damped exponentially in amplitude. Inductive coupling between loops and capacitive coupling between disks in such a hollow tube may be thought of as taking place in this way. The attenuation in decibels is then linear with displacement and follows the law given in the figure for the two cases of loop and disk coupling respectively. The loop-coupled attenuator has the advantage that its rate of attenuation is lower than that of any other mode and is thus less susceptible to errors due to undesired coupling to other modes, since these die away more rapidly and become unimportant at any but the lowest attenuator settings.

At low frequencies, a coil of several turns is usually employed in order to reduce the minimum insertion loss. For low-frequency applications where the skin depth is appreciable, this effect must be taken into account since it modifies the effective diameter. This may be done by adding $2p$ to the measured diameter, where p is the skin depth. For copper, p is given by:

$$p = \frac{6.6}{\sqrt{f}} \times 10^{-3} \quad \text{cm} \tag{15}$$

where f = megacycles per second.

The input and output impedance of this type of attenuator is reactive, and it is necessary to provide resistive pads to match the generator and load. These pads contribute to the initial insertion loss, which is of the order of 25 db in practical designs. The high initial insertion loss limits the use of this type of attenuator, although it is satisfactory for many signal generator applications where the power level of an oscillator must be reduced to the noise level of a sensitive receiver.

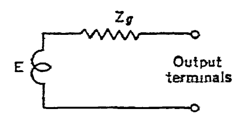


FIG. 23. Equivalent Circuit of Attenuator Output

The power delivered to a load impedance from any attenuator system can be analyzed in terms of Thévenin's theorem as shown in Fig. 23. Here z_g is the impedance seen looking into the output terminals and E is the open-circuit voltage measured at these terminals. At low frequencies, z_g is often made small compared to the load impedances with which the generator will operate so that the output voltage is essentially independent of z_L. A dummy antenna which simulates the impedance of an actual receiving antenna is then placed between the signal generator terminals and the receiver input.

At higher frequencies, line length between the signal generator and load cause considerable difficulty so that a better scheme is to make z_g equal to the characteristic impedance of the line which will be used. The load impedance is then also matched to the line so that the voltage at the load is $E/2$. If the receiver to be tested does not present an input impedance equal to the characteristic impedance of the line, it may be transformed to this impedance by means of a suitable transformer. If this is not done, an error will be introduced which can be calculated by means of the transmission-line equations.

40. SHIELDING PROBLEMS

There is usually a difference in level of the order of 100 db between the signal generator oscillator power and the attenuator output power when the generator is used for receiver testing. This makes necessary extreme precautions in the matter of shielding. The depth of penetration in a metal is defined as that depth to which the electric or magnetic field falls off to $1/\varepsilon$ of its value at the surface, where $\varepsilon = 2.718$. An attenuation of 100 db requires 11.5 skin depths. At a frequency of 1 megacycle per second, this corresponds to 0.076 cm in copper. Thus, except for very low frequencies, the thickness of the metal required is determined by mechanical rather than electrical considerations. The real problem is concerned with the bringing in of d-c leads and control shafts to the oscillator and with the necessity of providing removable covers for the shield case. Separate shielding boxes should be provided for the main units in the signal generator such as the oscillator, the power measurer, and the attenuator system. These individual shields are then enclosed in an overall shield and are preferably grounded to it at a single point to eliminate circulating currents which can induce voltages in lead filters or other output connections. Gaskets made of compressed woven metal are useful in preventing leakage from removable covers.

For low-frequency lead filters, conventional LC-type filters may be used. At higher frequencies, particular attention must be paid to the condensers used since these may, in fact, become inductive. Condensers of the button type are particularly useful for this purpose.

At frequencies above 2000 megacycles per second, lossy filters may be employed to advantage. A filter consisting of a coaxial line having powdered iron between the conductors represents one such type of filter which has seen considerable application. At 3000 megacycles per second, such a filter, 4 in. in length, can be made to have an attenuation of more than 100 db.

Control shafts may conveniently be brought into shielded generators by utilizing the wave-guide-beyond-cutoff principle and employing Bakelite shafts feeding through metal tubes which are soldered to the shielding box. A tube having an inside diameter of $^1/_4$ in. will, for instance, have an attenuation of about 100 db for each $^3/_4$ in. of length for frequencies up to 15,000 megacycles per second. Another method of feeding shafts through a shielded partition is by means of a flexible diaphragm type of coupler which solders to the shield box and which transforms a nutating motion into a rotary one in a manner which allows metallic continuity through the coupling.

41. POWER MEASUREMENT

Power measurers are of two types, one which samples the power in a transmission line between generator and load and which absorbs only a small fraction of the power being measured, and a second type which absorbs all the power being measured, in general converting it to heat. A vacuum-tube voltmeter of impedance high compared to the generator or load is an example of the former type; a bolometer or crystal power measurer terminating a transmission line is an example of the latter type.

All the types of power measurers discussed under the section on signal generators may be extended to measure higher powers through the use of attenuators appropriate to the frequency range involved. Bolometer-type power measures of the hot-wire or thermistor type are particularly adaptable to this use since they may conveniently be matched into coaxial lines. Again, broad-band designs may be evolved up to frequencies where the length of the bolometer element becomes an appreciable part of a wavelength. By means of special matching techniques, broad-band designs have been made up to frequencies of 10,000 megacycles per second.

Attenuators of resistive center conductor coaxial line using the thin metallic film on glass techniques can measure powers as great as a few watts. For higher power, lossy attenuators which can be matched over a considerable frequency range can be made by filling the space between conductors in a coaxial line with a lossy compound such as may be made with graphite or iron powder and a suitable cement binder. With proper proportions of lossy material to binder the loss per unit length can be adjusted to permit the power absorbed to be radiated adequately from the length of line used.

BIBLIOGRAPHY

Terman, F. E., *Radio Engineering*, McGraw-Hill Book Co., 1938.
Terman, F. E., *Measurements in Radio Engineering*, McGraw-Hill Book Co., 1939.
Sarbacher, R. I., and Edson, W. A., *Hyper and Ultra-high Frequency Engineering*, John Wiley and Sons, 1943.
McArthur, E. D., and Spitzer, E. E., Vacuum Tubes as High Frequency Oscillators, *Proc. I.R.E.*, Vol. 19, No. 11, 1971 (November 1931).
Seeley, S. W., and Anderson, E. I., U.H.F. Oscillator Frequency Stability, *R.C.A. Rev.*, Vol. 5, 77 (July 1940).
Harrison, A. E., *Klystron Technical Manual*, Sperry Gyroscope Co., 1944.
Karplus, E., The Butterfly Circuit, *General Radio Experimenter*, Vol. 19, No. 5 (October 1944).
Mayer, H. F., Visual Alignment Generator, *Electronics*, Vol. 13, No. 4, 39 (April 1940).
Crosley, M. G., Reactance Tube Frequency Modulator, *Q.S.T.*, June, 1940.
Sheaffer, C. F., Frequency Modulator, *Proc. I.R.E.*, Vol. 28, No. 2, 66 (February 1940).
Peterson, A., Vacuum Tube and Crystal Rectifiers as Galvanometers and Voltmeters at Ultra-high Frequencies, *General Radio Experimeter*, Vol. 19, No. 12 (May 1945).
Becker, J. A., Green, C. B., and Pearson, G. L., Properties and Uses of Thermistors—Thermally Sensitive Resistors, *Electrical Engineering*, Vol. 65, No. 11, *Transactions*, p. 711 (November 1946).
Smith, M. T., An Improved Ultra-high Frequency Signal Generator, *General Radio Experimenter*, Vol. 15, No. 8 (February 1941).
Scott, H. H., Progress in Signal Generator Design, *General Radio Experimenter*, Vol. 17, No. 6 (November 1942).
Harnett, D. E., and Case, N. P., The Design and Testing of Multirange Receivers, *Proc. I.R.E.*, Vol. 23, No. 6, 578 (June 1935).
McElroy, P. K., Designing Resistive Attenuating Networks, *Proc. I.R.E.*, Vol. 28, No. 2, 66 (February 1940).
Peterson, A., Output Systems of Signal Generators, *General Radio Experimenter*, Vol. 21, No. 1 (June 1946).
Shive, S. L., Effectiveness of Conduit as R.F. Shielding, *Electronics*, Vol. 19, No. 2, 160 (February 1946).

SECTION 12

ACOUSTICS

ACOUSTICS

THE SENSE OF HEARING

By John C. Steinberg and W. A. Munson

In the design of sound transmission and reproduction systems, consideration should be given to the physical and physiological properties of the voice and ear, and to the characteristics and psychophysiological effect of the different types of sounds that the systems are called upon to transmit and reproduce.

1. DESCRIPTION OF THE EAR

The hearing mechanism is usually divided into three parts, the outer ear, the middle ear, and the inner ear. A cross-section of the ear is shown schematically in Fig. 1.

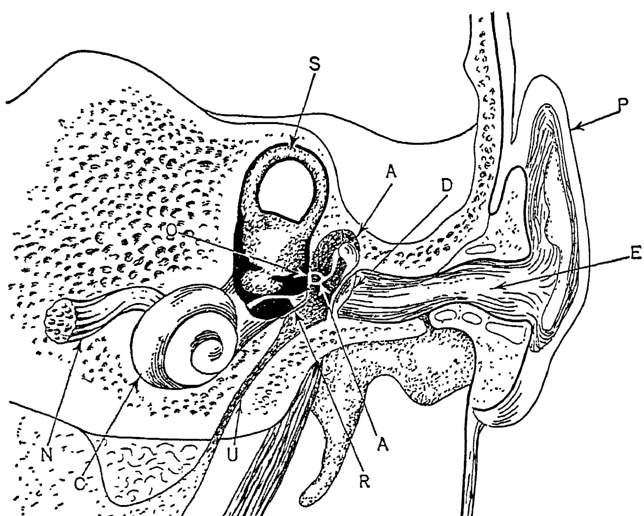


FIG. 1. Semi-schematic Section of Left Ear. *P*, pinna; *E*, ear canal; *D*, eardrum; *A*, auditory ossicles; *O*, oval window; *R*, round window; *U*, Eustachian tube; *S*, semicircular canal; *N*, auditory nerve; *C*, cochlea.

THE OUTER EAR. The ear canal (*E*) has a length of about 2.5 cm, a volume of 1.0 cu cm, and an area at the opening of 0.3 to 0.5 sq cm. The eardrum (*D*), stretched across the inner end of the canal, has a horizontal diameter of 1.0 cm, a vertical diameter of 0.8 cm, and an area of 0.6 sq cm. All these dimensions vary somewhat for different individuals.

THE MIDDLE EAR. The middle ear is separated from the outer ear canal by the eardrum. Motion of the eardrum is transmitted across the middle-ear cavity to the inner ear by means of a system of levers (*A*) called the auditory ossicles. The ossicles consist of three small bones, the malleus (hammer), attached to the eardrum, the stapes (stirrup), attached to the oval window (*O*) of the inner ear, and the incus (anvil), which is the connecting link between the malleus and stapes. The weights of the three bones are: hammer, 0.023 gram; anvil, 0.025 gram; and stirrup, 0.003 gram. The middle-ear cavity is normally filled with air maintained at atmospheric pressure by virtue of the Eustachian tube (*U*) which leads to the back part of the throat.

THE INNER EAR. The inner ear, which serves a dual purpose, is a complex labyrinth of liquid-filled passageways imbedded in the temporal bone adjacent to the middle-ear cavity. There are three semicircular canals, oriented so as to lie in three mutually perpendicular planes, and these ducts contain the nerve endings concerned with the maintenance of body equilibrium. Only one canal (*S*) is shown in Fig. 1. The cochlea (*C*) is a spiral duct containing the auditory nerve endings. In Fig. 1, it has been drawn on an enlarged scale relative to the middle and outer ears, and the bone in which the cochlea is imbedded is not shown. The mean diameter of a semicircular canal is about 1.0 cm. The cochlear spiral has a mean diameter of about 0.6 cm for the large turn at the base. As may be seen in the sectional view of Fig. 2, the cochlea is divided by flexible membranes into three canals, scala vestibuli (*V*), scala media (*M*), and scala tympani (*T*).

The auditory nerve fibers running between the cochlea and the brain number about 29,000 They terminate with complex interconnections along the basilar membrane (*B*),

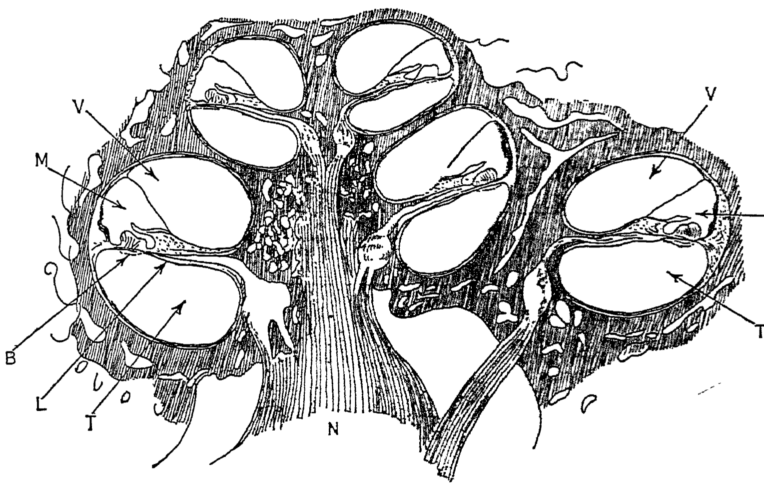


FIG. 2. Semi-diagrammatic Section of the Cochlea. (From Gray's *Anatomy*.) *B*, basilar membrane; *V*, scala vestibuli; *M*, scala media; *T*, scala tympani; *N*, cochlear nerve; *L*, lamina spiralis ossea.

which extends from the bony ledge (*L*) to the opposite wall of the cochlea, separating scala tympani and scala media. The nerve endings are associated with tiny hair cells which protrude into the liquid filling the small triangular scala media. Scala vestibuli and scala tympani are interconnected at the apex of the spiral by a small opening called the helicotrema. At the base of the spiral, scala tympani opens into the middle-ear cavity through the round window (*R*), shown in Fig. 1, but the liquid content is retained by a flexible membrane stretched across the opening. The oval window (*O*), between the middle ear and scala vestibuli, is closed by the foot of the stapes and connecting ligaments.

Excitation of the Auditory Nerves. A sound wave in the outer ear canal produces motion of the eardrum and associated ossicles which, in turn, agitate the liquid in scala vestibuli through the oval window. When the stimulus is a pure tone at a low level, this results in an excitation of a small number of the nerve endings enclosed by the adjacent scala media. When the frequency of the tone is changed, the excitation moves along the basilar membrane, engaging a different set of nerve endings. In this way the ear differentiates between tones of different pitch. When the intensity of the tone is increased, the excitation spreads out, and additional nerve endings are stimulated, but those comprising the initial group probably receive the maximum stimulation. The positions of the stimulation maxima are shown in Fig. 3 as a function of frequency.

At low frequencies the stimulation maxima are very broad and the positions are poorly defined. It may be that the pitch of low-frequency tones is sensed by the stimulation frequency as well as by its position.

NERVE CONDUCTION. Physiological research indicates that a nerve conducts on an "all-or-none" basis. The magnitude of the nerve impulse and the rate of propagation are independent of the manner of excitation. After an impulse has been transmitted by a nerve, it must recover its conducting properties before it is capable of being excited again. This recovery period lasts for a time interval of 1 to 3 milliseconds and varies somewhat with the intensity of the stimulus. Thus a nerve fiber is unable to transmit the wave form of the exciting stimulus. Excitation of the auditory nerves and conduction of neural pulses to the brain have been studied by placing a small electrode on a single auditory nerve tract of anesthetized animals and recording the potential generated. It was found that increasing the intensity of a pure tone increases the rate at which pulses are transmitted. Typical results are shown in Fig. 4. Since a maximum rate is attained when the sound level is increased about 50 db, other fibers must be involved in covering the full intensity range of the ear. It is believed that the loudness of a sound is related to the rate at which neural impulses reach the brain from all portions of the basilar membrane.

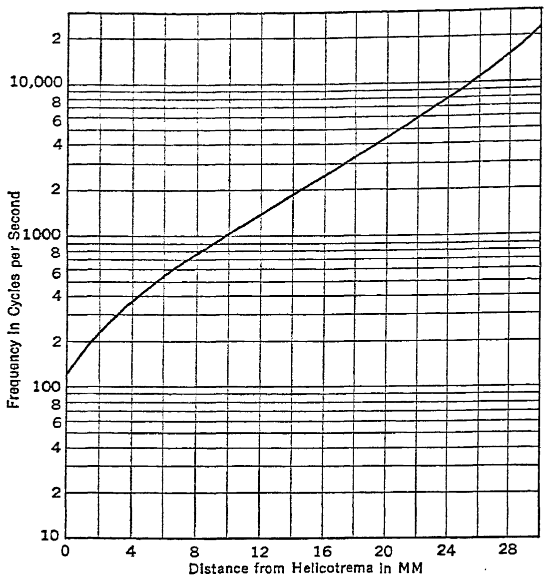


FIG. 3. Relation between Frequency of a Tone and the Position of Nerve Endings on the Basilar Membrane Which Receive Maximum Stimulation. (Steinberg.)

ACOUSTIC IMPEDANCE OF THE EAR. The acoustic impedance which an ear presents to a telephone receiver has been measured at the aperture of a receiver cap. The results depend to a large extent upon the way the cap fits the ear. Typical values are shown in Fig. 5 for the case when the cap is sealed to the ear and for the case when an air leak is present between the cap and the ear. These data are useful in receiver design, and Inglis, Gray, and Jenkins describe an artificial ear which they use for the measurement of telephone receivers. It consists of a conduit having the approximate dimensions and impedance of a typical ear canal over the important frequency range. The exposed end of the conduit is fitted with a tapered rubber seating surface on which the receiver is placed. The other end of the conduit is terminated by an acoustic network and by the diaphragm of a small condenser transmitter. The pressures developed at the transmitter diaphragm for a given voltage on a receiver placed on the artificial ear are closely equal to the pressures that would be produced by the receiver at the drum of a typical human ear.

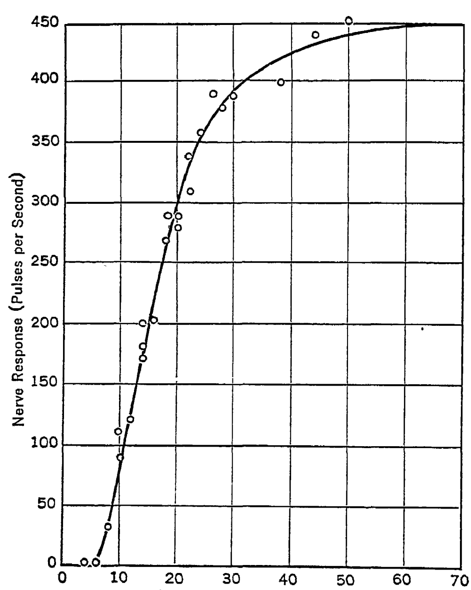


FIG. 4. Relation between Sound Intensity and Nerve Discharge Rate for a 1050-cycle Tone. (Galambos and Davis.)

NATURAL FREQUENCY AND DAMPING CONSTANT OF THE EAR. The work of Békésy indicates that the frequency with which the eardrum and ossicles vibrate when suddenly released from a displaced position is of the order of 1200 to 1500 cycles. The vibrations decay at a rate of 1200 db per sec. Davis and associates' experiments on the electrical response of the ears of cats indicate a natural frequency from 600 to 1000 cycles and a decay rate of 1000 db per sec. Their work also indicates that the time lag between the sudden application of a sound wave on the eardrum and the transmission of the impulse along the auditory nerve is of the order of 2 to 3 milliseconds. On the other hand, the sensory build-up time, e.g., the time needed for a suddenly applied steady wave to build up to a steady loudness, is of the order of 0.2 to 0.25 sec.

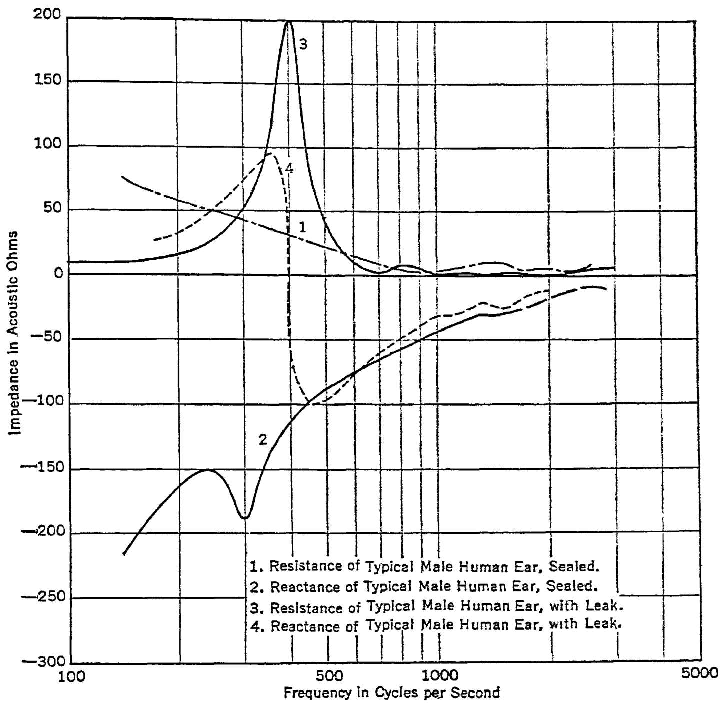


FIG. 5. Acoustic Impedance of Ears as Viewed through Aperture of Receiver Cap. (Inglis, Gray, and Jenkins.)

2. SENSITIVITY OF THE EAR

Ear sensitivity is concerned with the least intense sound that can be heard. Such a sound is said to be at the "threshold of hearing." Two classes of ear-sensitivity determinations have been reported, "minimum audible pressure" (M.A.P.) and "minimum audible field" (M.A.F.). M.A.P. is the just-audible sound pressure measured near the observer's eardrum. M.A.F. is the free field sound intensity of a plane progressive wave that is just audible to an observer facing the source and listening binaurally. The sound intensity is measured before the observer's head is inserted in the field. It is convenient to express minimum audible values in decibels from an arbitrary reference. Reference intensity has been chosen as 10^{-16} watt per square centimeter. Corresponding reference pressure at 20 deg cent and 76 cm of Hg is 2×10^{-4} dyne per square centimeter. Intensity and pressure levels are the number of decibels from the above references, respectively. Table 1 gives minimum audible values derived by Sivian and White from their own work and that of others when a small selected group of observers with excellent hearing is used.

Owing to the diffraction effect of the head and to the difference between one-ear and two-ear listening, it would be expected that M.A.P. and M.A.F. would differ considerably at high frequencies but not at low frequencies. Various possible causes of the differences shown in Table 1 at the low frequencies are discussed by Sivian and White, with the conclusion that, at the present time, a satisfactory explanation is not evident.

Table 1. Monaural Minimum Audible Pressure Levels in the Ear Canal and Binaural Minimum Audible Sound Field Intensity Levels

Frequency	60	100	200	500	1,000	2,000	5,000	10,000	15,000
M.A.P., db (Monaural)...	59	46	29	14	8	5	12	26	44
M.A.F., db (Binaural)....	45	33	19	8	3	−6	−3	11	22

BINAURAL VS. MONAURAL. Minimum audible values obtained with observers listening binaurally (two ears) are smaller on the average than those obtained with observers listening monaurally (one ear). According to Fletcher and Munson, the difference appears to be accounted for by inequalities in the sensitivities of the two ears of an observer. They report that the binaural sensitivity is practically equal to that of the better ear, and that monaural sensitivity is equal to that obtained by averaging the sensitivities of both ears. Figure 6 gives the differences between the sensitivity of the better

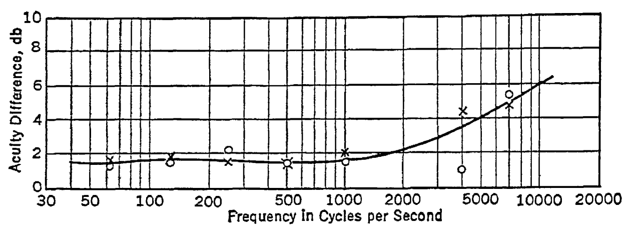


FIG. 6. Difference between the Sensitivity of the Better Ear and the Average of Both Ears. (Fletcher and Munson.)

ear and the average of both ears, based on the audiograms of 80 persons of normal hearing. It gives a means of converting binaural M.A.F. into monaural M.A.F.

VARIATION WITH DIRECTION. The values of M.A.F. given in Table 1 are for an observer facing the source, i.e., for a progressive wave whose wave front is vertical and whose direction is normal to a line joining the observer's ears. Owing to diffraction effects of the head, the ear is directive and generally hears best when the open ear is turned toward the source. The directivity of hearing as reported by Sivian and White is given in Fig. 7. Directivity is expressed as the variation in monaural M.A.F. with the azimuth of the vertical wave front: 0° corresponds to the observer facing the source; +90° corresponds to the open ear toward the source. The variation is expressed in decibels from the M.A.F. at 0°. A positive value of directivity for any angle means that the ear is more sensitive at that angle than at 0°. By means of the directivity data, Sivian and White have computed the sensitivity of the ear for the case when an observer is exposed to a diffuse sound field, such that sound waves of equal amplitudes and random phase angles are equally probable from all angles. Their results are shown in Table 2.

Table 2. Binaural Minimum Audible Sound Field (Random Horizontal Incidence).
(Sivian and White)

Frequency	60	100	200	500	1,000	2,000	5,000	10,000	15,000
Binaural M.A.F., db (random horizontal incidence)............	45	33	19	6	−1	−7	−7	2	18

POSSIBLE LOWER LIMITS OF SENSITIVITY. On certain assumptions, Sivian and White calculate that the intensity level of thermal-acoustic noise, i.e., noise originating from the thermal velocities of air molecules, has a value of 11 db below 10^{-16} watt for a frequency band from 1000 to 6000 cycles, and hence is of the order of the maximum ear sensitivity. Although the authors regard the calculation as very approximate, it suggests that the limit of sensitivity may be set by the transmitting medium. If this is so, man may have maximum sensitivities comparable with those of animals.

EAR SENSITIVITY OF THE POPULATION. The values of ear sensitivity shown in Table 1 are for a selected group of young people with excellent hearing; not all people hear this well. During the New York and San Francisco World Fairs in 1939, which were attended by people from all parts of the country, a survey of hearing was made in a large group of the United States population. The results, which are plotted in Fig. 8, show the percentages of thresholds measured which were above the levels shown by the contour lines. Thus the data show the percentage of people in the population within the age range from 10 to 59 years who cannot hear tones below the levels indicated by the contours.

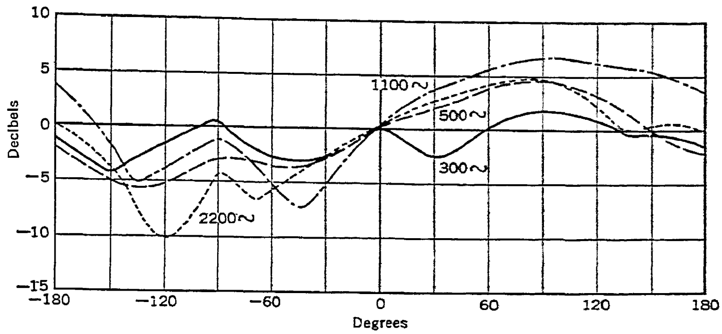


FIG. 7a. Increase of Monaural Acuity in a Free Sound Field When the Direction Which the Observer Faces Is Changed

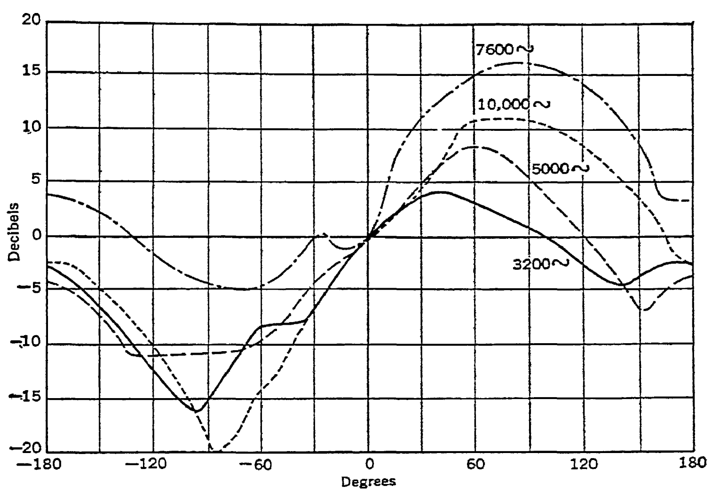


FIG. 7b. Increase of Monaural Acuity in a Free Sound Field When the Direction Which the Observer Faces Is Changed

For instance, 25 per cent of the population cannot hear a 1000-cycle tone if the intensity is below the 20-db level (zero = 10^{-16} watt per sq cm). The dotted portions represent extrapolations of the distributions beyond the intensity and frequency ranges used in the tests and are, of course, speculative.

VARIATION WITH AGE. Further analysis of the data revealed a progressive deterioration of hearing with age for the average individual, particularly at the higher frequencies. The hearing loss is somewhat greater for men than for women, as shown by the curves in Fig. 9. For this plot "zero hearing loss" refers to an ear sensitivity equal to the median value for the population tested. The contour marked "50 per cent" in Fig. 8 shows the median values as a function of frequency. When the hearing loss exceeds about 25 db at all frequencies lower than 2000 cycles, a person will experience difficulty at times in understanding speech in classrooms, auditoriums, churches, and similar environments where the speech level is not very high. A hearing loss of 45 db for frequencies below 2000 cycles results in some difficulty in understanding speech at distances greater than

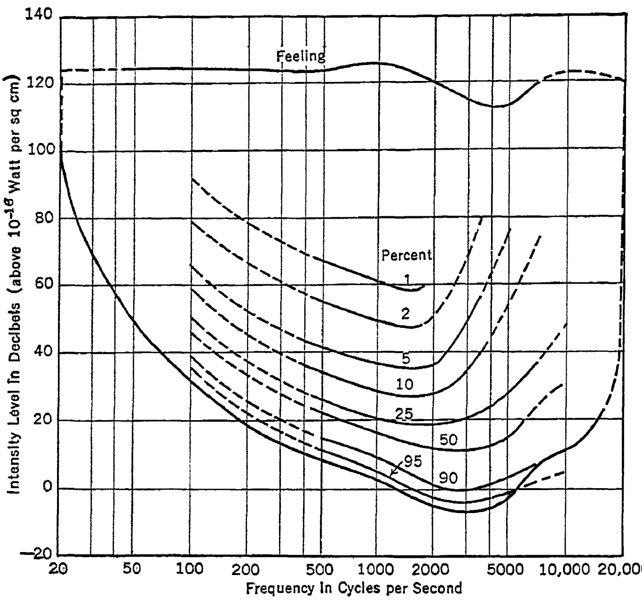


FIG. 8. Contour Lines Showing the Lower Limit of the Sensitivity for a Given Percentage of Tests in the Age Group 10–59. (Steinberg, Montgomery, and Gardner.)

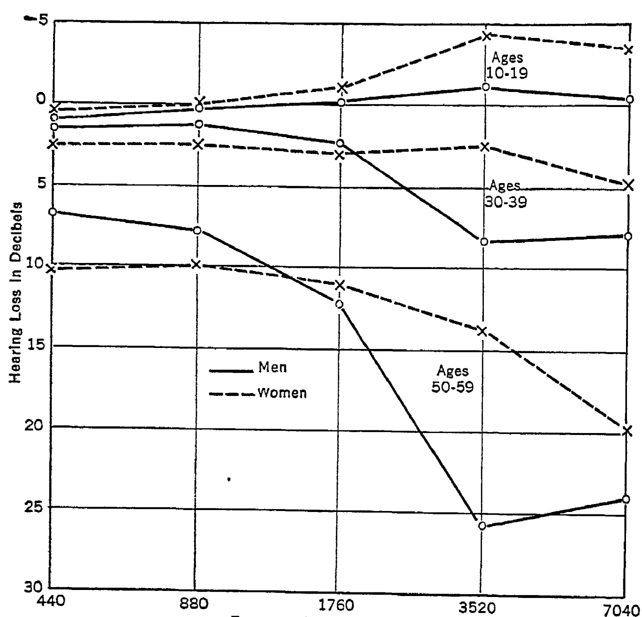


FIG. 9. Average Hearing Loss for Men and Women in Three Age Groups. (Steinberg, Montgomery, and Gardner.)

2 or 3 ft, and considerable benefit is usually derived from the use of a hearing aid. Table 3 shows the prevalence of hearing losses greater than 25 and 45 db among the people tested.

In the age group from 50 to 59 years, Table 3 shows that 3 or 4 per cent of the population have hearing losses exceeding 45 db and, therefore, would experience some difficulty understanding speech at a distance of 2 or 3 ft. Until the hearing loss exceeds about 65 db, a person will still be able to use the telephone without much difficulty.

Table 3. Percentage of Tests with Hearing Losses Greater than 25 db and 45 db

(Steinberg, Montgomery, and Gardner)

Age Group	25-db Loss				45-db Loss		
	Frequency				Frequency		
	880	1760	3520	7040	880	1760	3520
10–19							
Men	1.7	1.6	4.5	8	0.6	0.6	1.8
Women	1.8	1.2	1.2	2.5	0.6	0.4	0.3
20–29							
Men	1.1	1.2	7	9.5	0.1	0.3	2.7
Women	1.8	1.6	2.2	3.5	0.4	0.3	0.7
30–39							
Men	1.8	3.5	15	19	0.3	0.6	6
Women	3.5	3.5	5.5	10	1.2	0.8	1.6
40–49							
Men	5.5	9.5	32	39	1.4	2.6	16
Women	7	7	11	24	2.1	1.5	3
50–59							
Men	9.5	17	48	58	2.6	6	27
Women	13	14	22	43	4	3	7

AUDITORY RANGE. As the intensity level of a sound is increased, a value is reached which causes a sensation of feeling in addition to the sensation of tone. Such intensity levels may be taken as practical upper limits of hearing, and they have been called the "threshold of feeling." The observed data on the threshold of feeling were the number of decibels between the hearing and feeling thresholds for pure tones and were obtained by means of telephone receivers held to the ear. The absolute threshold of feeling plotted in Fig. 8 was obtained by adding the observed data to the lower limit of audibility shown in Fig. 8. The lower limit is the M.A.F. determination of Table 1.

The lower limit and the feeling curve have been extrapolated until they meet, thus forming an enclosure called the "auditory sensation area." This area has the property that any sinusoidal sound wave having a frequency and an intensity level within the area will cause a sensation of tone. The dashed portions of the curves serve as a means of defining the upper and lower frequency limits, i.e., the lowest and the highest frequency that can be sensed as a tone. Measurements have been made using frequencies varying from about 8 to 40 cycles for the lower limit and from 12,000 to 35,000 cycles for the upper limit, but, for the most part, very little attention was given by the experimenters to the intensity levels at which the determinations were made. In drawing in the dashed curves, consideration was given to available data on frequency limits, and values of 20 cycles and 20,000 cycles were selected for the lower and upper frequency limits, respectively. These limits were taken as more or less typical for persons of normal hearing in the age range from 18 to 23 years but do not represent the extreme values found in exceptional cases.

3. DIFFERENTIAL SENSITIVITY

The differential sensitivity of hearing refers to some aspect of the smallest change of a stimulus that can be detected. Available data on the subject are confined almost entirely to measurements of the minimum perceptible increments of frequency and of intensity when the stimulus is a single-frequency tone. The values obtained depend to some extent upon the method of measurement, probably because the just perceptible change (difference limen) is affected by the rate at which the change is made as well as other variables common to tests of this type. Until a measuring technique is standardized, the data shown should be regarded as of the exploratory type. In Fig. 10 the results of extensive measurements of Δf, the difference limen for frequency, are plotted. For levels greater than 40 db above threshold, and for frequencies greater than 500 cycles, the frequency D.L. has the approximately constant value of 0.3 per cent.

These results were obtained by listening to the tones with head receivers. When the measurements are made by listening to a loudspeaker in a room, much smaller values can be observed. In this case the change in frequency is detected by intensity changes due to the shifting of the interference pattern in the room with frequency change. A

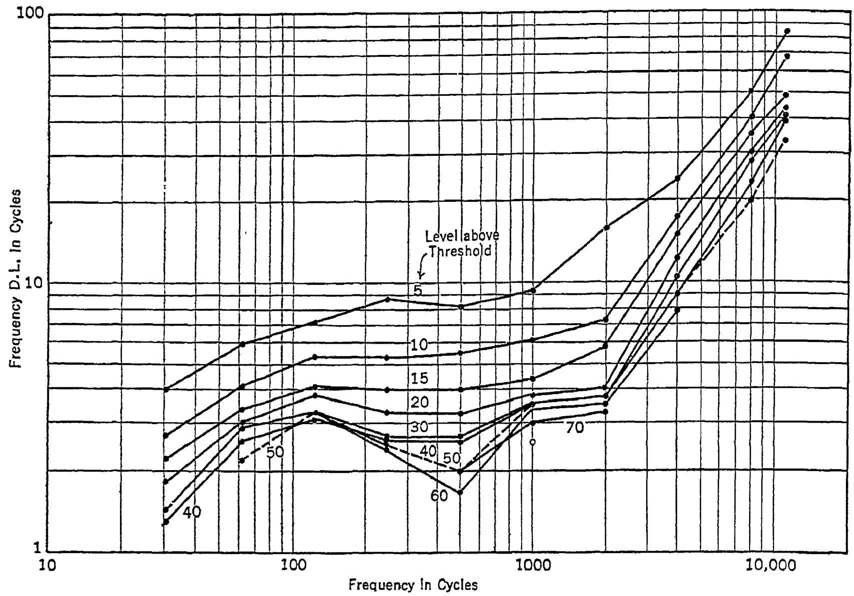


FIG. 10. The Frequency Difference Limen with Decibels above Threshold as a Parameter. (Shower and Biddulph.)

similar set of measurements of ΔI, the minimum perceptible increment of intensity, is shown in Fig. 11. For convenient use the increment of intensity expressed in decibels, $10 \log (I + \Delta I/I)$, is plotted as the ordinate. For levels greater than 40 db above threshold, and for frequencies between 200 and 7000 cycles, the intensity D.L. varies from 0.25 to 0.75 db.

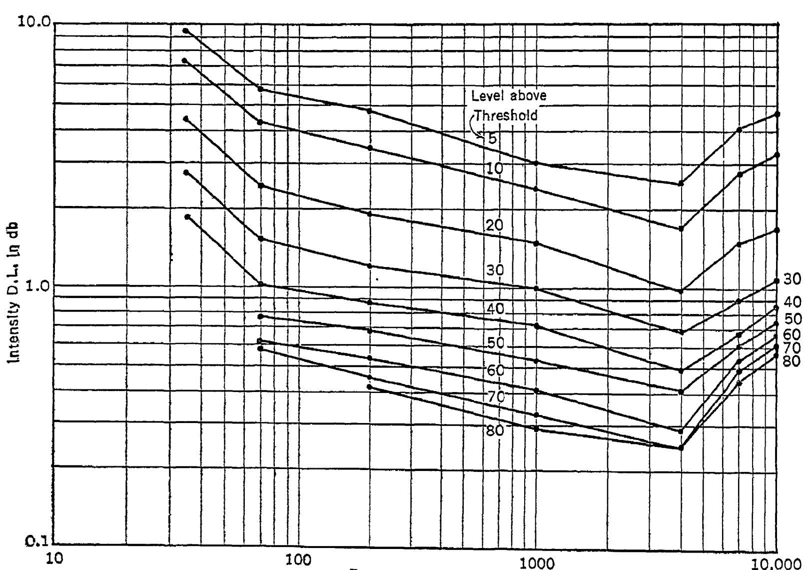


FIG. 11. The Intensity Difference Limen with Decibels above Threshold as a Parameter. (Riesz.)

4. MASKING EFFECTS OF SOUNDS

When listening to speech or music it often occurs that a disturbing noise interferes to such an extent that the desired sounds are partially or entirely obliterated. The noise is said to have a masking effect, and the magnitude of the effect is defined by a "masking spectrum" of the noise. Measurements of masking spectrums are made by determining the threshold of audibility of single-frequency tones in the presence of the noise and again when the noise is absent. If β is the intensity level of a tone of frequency f, that is just audible in the presence of a noise, and β_0 is the intensity level that is just audible under quiet listening conditions, then M, the masking at the frequency f, is defined by the equation: $M = \beta - \beta_0$. A plot of M as a function of the frequency of the tones is called a masking spectrum. If the intensity spectrum of a noise has been measured by means of a sound meter and filters, the masking spectrum may be derived by use of Fig. 12. Here the iso-masking intensity per cycle level of noise has been plotted as a function of frequency with masking as a parameter. The applications of these data are limited to portions of the noise intensity spectrum that do not exhibit the abrupt changes with frequency that occur when a prominent single-frequency component is present or a filter with a sharp cutoff limits the frequency range of the noise.

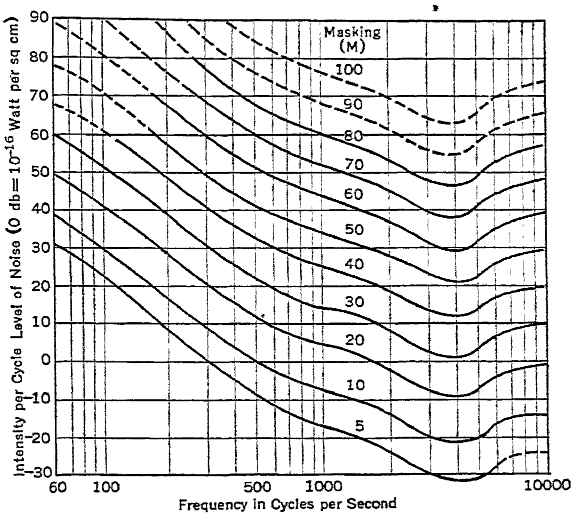


FIG. 12. Masking Contours for a Steady Noise. (Fletcher and Munson.)

5. LOUDNESS OF SOUNDS

Loudness is defined as the magnitude of an auditory sensation, and for steady sounds it is thought to be proportional to the rate at which neural pulses originating along the basilar membrane arrive at the brain. A scale of auditory magnitudes has been derived from loudness tests and can be used whenever the loudness level of a sound is known. A measurement of loudness level consists of a listening test in which the level of a 1000-cycle reference tone is adjusted until it sounds equally loud to the sound being measured. Errors in judgment may be large, and a number of observers may be required to obtain a reliable average result. The equivalent free field intensity level of the equally loud 1000-cycle tone is, by definition, the loudness level of the unknown sound.

The loudness level of a sound being known, its auditory magnitude is found by referring to Table 4, which gives the auditory magnitude, or loudness, as a function of the intensity level of the equally loud 1000-cycle tone. The figures for the intensity level (loudness level) of the 1000-cycle tone are in decibels relative to 10^{-16} watt per sq cm, and a level of 40 db results in an auditory magnitude of one sone. Thus, in Table 4, the loudness corresponding to a loudness level of 40 db is 1 loudness unit (1 sone = 1000 millisones). Increasing the loudness level from 40 db to 49 db gives a listener the impression that the magnitude has doubled, so a loudness level of 49 db results in a loudness of 2 units. Increasing the loudness level from 49 to 58.1 db again doubles the magnitude of the sensation for the average listener, so a loudness level of 58.1 db produces a loudness of 4 units. The empirical relationship between loudness and loudness level is shown in Table 4.

Loudness levels are given in steps of 1 db, or 1 "phon," and zero level is 10^{-16} watt per sq cm in a free sound field. The term "phon" is generally used as the unit of loudness level to avoid confusion with intensity levels of sounds other than a 1000-cycle tone. It is believed that the loudness of a steady sound depends upon the rate at which neural pulses reach the brain, and, although no measurements are available to substantiate this hypothesis, the concept is useful in explaining empirical methods of computing the loudness of complex sounds. It follows from the manner in which the loudness scale was developed that a 1-sone sound, for example, is equally loud to any other 1-sone sound, and four times as loud as a 0.250-sone sound. It is also true that a 1-sone sound will drop to a

Table 4. Loudness (Millisones) vs. Loudness Level (Phons)

(Fletcher and Munson)

Loudness Level, db	Loudness 0	1	2	3	4	5	6	7	8	9
0	1.0	1.42	1.95	2.58	3.36	4.32	5.57	7.10	9.00	11.4
10	14.4	18.7	23.3	28.9	35.1	42.2	50.6	60.3	71.6	85.0
20	100	120	142	165	188	214	242	272	307	340
30	380	421	470	522	577	635	700	763	835	915
40	1,000	1,080	1,170	1,260	1,360	1,470	1,590	1,710	1,850	2,000
50	2,150	2,330	2,510	2,710	2,930	3,160	3,410	3,690	3,980	4,300
60	4,640	5,010	5,410	5,840	6,310	6,810	7,360	7,940	8,580	9,260
70	10,000	10,800	11,700	12,600	13,600	14,700	15,900	17,100	18,500	20,000
80	21,500	23,300	25,100	27,100	29,300	31,600	34,100	36,900	39,800	43,000
90	46,400	50,100	54,100	58,400	63,100	68,100	73,600	79,400	85,800	92,600
100	100,000	108,000	117,000	126,000	136,000	147,000	159,000	171,000	185,000	200,000

half sone if only one ear is used for listening. These simple relations have all been verified experimentally. A complex sound having two 1-sone components will be equally loud to a sound of 2 sones if the components excite nerve endings located in different sections of the basilar membrane and thus contribute 1 sone each, even though they are sounded simultaneously. This is not likely to be the case when the frequencies of the components are close together or the levels are high, since many of the same nerve endings are then used by both components.

EFFECTIVE STIMULATION DENSITY. The response of the auditory nerve endings is dependent upon the density of stimulation. If the stimulus is localized, the response will be different, and the loudness sensation will differ, in general, from the loudness resulting when the stimulus energy is distributed among a large number of nerve endings. The relationship between the frequency of the stimulus and the position coordinate (X) of the nerve endings stimulated is shown in Fig. 13.

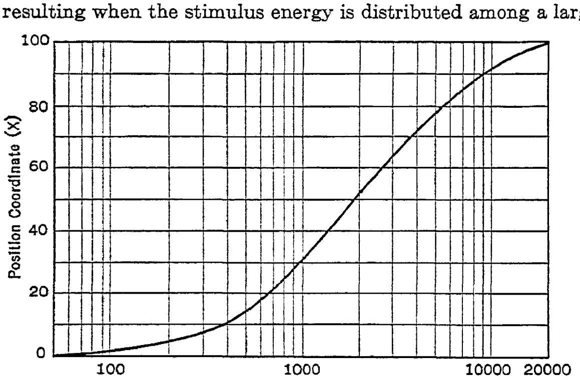


FIG. 13. Relation between Frequency of the Stimulus and the Position of Maximum Stimulation. (Fletcher.)

The ordinate of Fig. 13 gives the position of maximum stimulation of the nerve endings, with respect to the total number, when the stimulus is a single-frequency tone. For instance, at 1000 cycles the curve shows that 31 per cent of the nerve endings are on one side of the point of maximum stimulation, and 69 per cent are on the other side. The relationship shown can be derived in several different ways on the basis of different assumptions, but it has been verified only in a qualitative sense by actual nerve counts. To obtain the effective stimulation density as a function of frequency, for a sound with a continuous energy spectrum, measurements must first be made with a sound meter that will analyze sound at all frequencies within the audible range. If the analyzer measures the intensity (I) in a frequency band Δf cycles wide, then the readings are related to effective stimulation densities by means of the equation:

$$Z = 10 \log \frac{I \, \Delta f_1}{I_1 \, \Delta f}$$

where I_1 is the intensity level of a 1-millisone single-frequency tone, and Δf_1 is a frequency band enclosing a unit group (1 per cent) of the nerve endings. It is seen that Z is the ratio, in decibels, of the stimulation intensity per unit group of nerve endings to the intensity required to excite a 1-millisone response from the nerves. The equation for Z is usually given in the more convenient form:

$$Z = B + \kappa - \beta_0$$

where B is the measured intensity per cycle level of noise $[B = 10 \log (I/I_0 \Delta f)]$, I_0 is the reference intensity of 10^{-16} watt per sq cm, κ is the band width in decibels which includes a unit nerve group ($\kappa = 10 \log \Delta f_1$), and β_0 is the intensity level of a tone when the loudness level is zero $[\beta_0 = 10 \log (I_1/I_0)]$. Values of κ, B_0, and X are shown in Table 5 for use in computing and plotting the effective stimulation density for the nerve endings as a function of their position coordinate, X.

Table 5. Values of κ, β_0, and X

(Fletcher and Munson)

f	100	200	300	500	700	1,000	2,000	3,000	5,000	7,000	10,000
X	1	4	7	14	21	31	52	64	78	85	92
κ	16.5	15.0	15.0	15.0	15.3	16.0	18.6	20.7	23.9	26.1	28.6
β_0	37.4	22.9	14.6	5.7	2.0	0.0	−4.5	−8.5	−4.9	4.9	8.8

LOUDNESS COMPUTATION FOR SOUNDS WITH CONTINUOUS ENERGY SPECTRUMS. The loudness of a steady sound is believed to be proportional to the number of nerve pulses arriving at the brain in unit time from all parts of the basilar membrane. We may compute the loudness (N) from the expression:

$$N = \int_0^{100} N_X \, dX$$

where N_X, the sone density, is a measure of the number of nerve pulses originating from a unit group of nerve endings in unit time, and X is the position coordinate previously defined. The equation is solved by plotting N_X as a function of X and measuring the area under the curve. Thus the loudness, N, of any sound can be computed when the sone density is known as a function of X.

The relationship between sone density (N_X) and the effective stimulation density (Z_X) has been investigated for sounds characterized by a continuous distribution of energy, and it is shown in Fig. 14.

It is not applicable to sounds having single-frequency components or sounds in which the value of the slope, dZ/dX, exceeds the limits ± 2 db. In the latter case, masking measurements are a more reliable means of obtaining the sone density (N_X). The masking (M) is defined by the equation: $M = \beta - \beta_0$, where β is the intensity level of a single-frequency tone that is just audible in the presence of a masking sound, and β_0 is the intensity level when the loudness level is zero. Figure 15 shows the relationship between

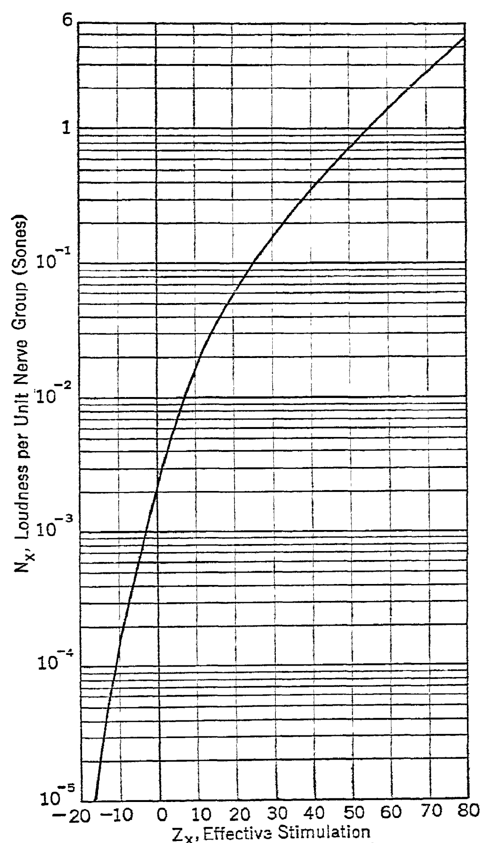


FIG. 14. Relation between the Loudness (N_X) per Unit Nerve Group and the Stimulation (Z_X) per Unit Nerve Group. (Fletcher and Munson.)

masking and sone density. The masking method of obtaining the sone density is valid for sounds exhibiting large values of dZ/dX but not for sounds with single-frequency components. The masking spectrum of a single-frequency tone is not an accurate indication of its sone density since the phenomenon of beats between the masking tone and the masked tone changes the conditions under which the masked tone is detected. This results in low values of masking at frequencies where the highest values would be expected.

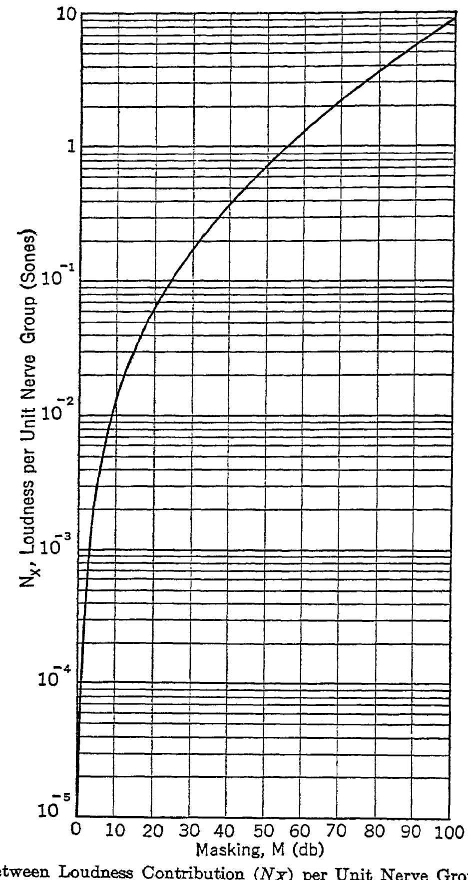


FIG. 15. Relation between Loudness Contribution (N_X) per Unit Nerve Group and Masking (M). (Fletcher and Munson.)

LOUDNESS OF SOUNDS COMPOSED OF SINGLE-FREQUENCY COMPONENTS. When a sound is a single-frequency tone at a low level, the nerve endings in a localized region of the basilar membrane are stimulated. As the level is increased, the stimulation pattern spreads to adjacent nerves, and at high levels a large proportion of all the nerves may come into use. The region of maximum stimulation probably does not change much when the level is raised, but secondary maxima appear at harmonic frequencies owing to non-linear distortion in the ear itself. The masking spectrums shown in Fig. 16 indicate the nature of the change of the stimulation pattern as the level of a tone is increased.

As explained in the previous section, the regions of maximum stimulation appear as depressions because of the beats which occur near the fundamental and harmonic frequencies. Measurements of the conditions for best beats have been used to determine the magnitude of harmonics produced in the ear. Figure 17 shows the equivalent magnitude of the subjective harmonics for different pressure levels of the applied sound. Taking the fundamental as the first harmonic, the abscissa gives the number of the harmonic. When plotted in terms of pressure level, the curves are independent of frequency. It is clear that the stimulation patterns of single-frequency tones are very complex, and it would be difficult to plot the sone density as a function of X, as was done for sounds with continuous energy spectrums. However, the loudness of any single component may be found from the loudness-level contours shown in Fig. 18.

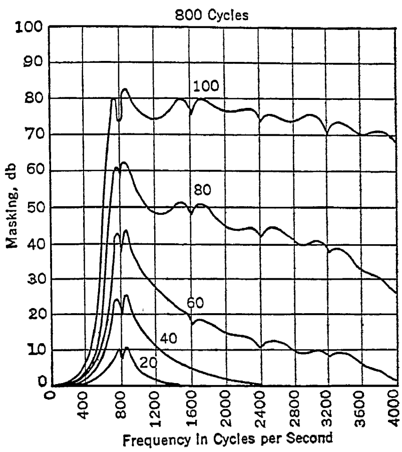


FIG. 16. Masking Spectrums of a Single-frequency Tone with Level above Threshold as a Parameter. (Fletcher, Wegel, and Lane.)

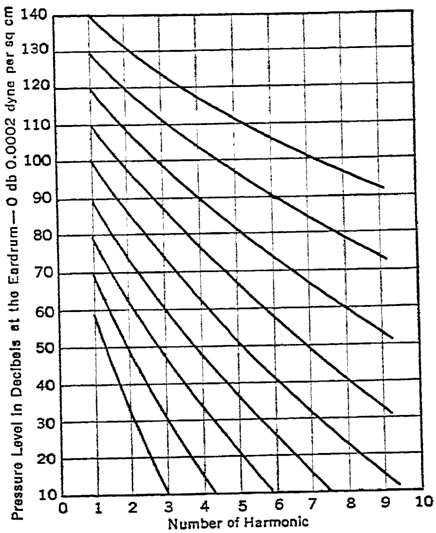


FIG. 17. Magnitude of Subjective Harmonics. (Fletcher and Graham.)

The ordinate here is the free field intensity level of a pure tone, and each contour line is marked with a loudness level. For instance, a 200-cycle tone at an intensity level of 60 db from 10^{-16} watt has a loudness level of 50 db. Turning to Table 4, we see that the corresponding loudness is 2.150 sones. If a sound has more than one component, the sone values may be added to obtain the total loudness, provided that the frequencies of the

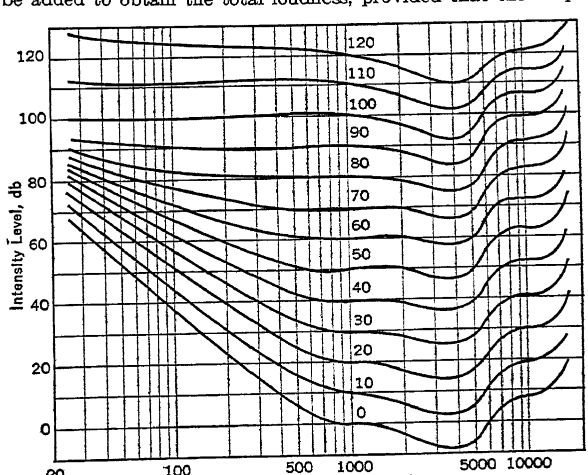


FIG. 18. Loudness-level Contours. (Fletcher and Munson.)

components differ enough so that different nerve groups are stimulated. The separation required is a complex function of level and frequency and will not be discussed here. Some idea of the effect of different separations of components may be obtained from

Fig. 19, which shows the loudness levels of several 10-component sounds as a function of the loudness level of each component.

The first sound had a fundamental frequency of 530 cycles and a difference of 530 cycles between components. The others had a fundamental of 1000 cycles and the differences indicated. The dotted line shows the loudness levels corresponding to 10 times the loudness of a single component. For a more extended treatment of the loudness of sounds with single-frequency components, refer to an article in the October 1933 *Journal of the Acoustical Society of America* entitled "Loudness, Its Definition, Measurement and Calculation," by Fletcher and Munson.

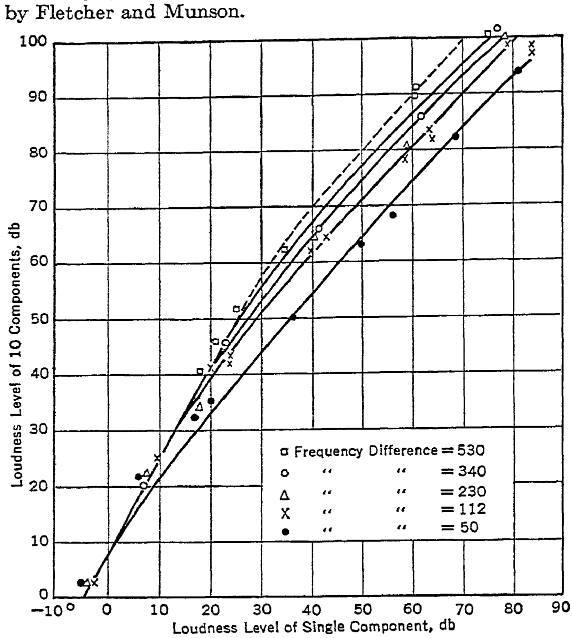


FIG. 19. Effect of Separation of Components on the Loudness Levels of Complex Tones. (Fletcher and Munson.)

6. THE PITCH OF STEADY SOUNDS

The pitch of a steady sound is the position on a musical scale that would be assigned to it by a listener. If the sound is a single-frequency tone, its pitch depends upon frequency and, to a slight degree, upon intensity. At loudness levels less than 40 db the pitch depends only upon frequency, and single-frequency tones at a loudness level of 40 db have been chosen as standards for comparison with sounds of unknown pitch. The results of pitch comparisons between single-frequency tones at a 40-db loudness level and tones at higher levels are shown in Fig. 20. The ordinate is the change in frequency, in per cent, of a tone that is necessary in order that its pitch remain constant as its level is raised.

For example, the curves show that a 100-cycle tone must be lowered 10 per cent when the loudness level is increased from the standard level to 100 db. This means that a 90-cycle tone at a loudness level of 100 will appear to have the same pitch as a 100-cycle tone at a loudness level of 40 db. The curves shown in Fig. 20 are based on data at low frequencies. Similar experiments by Zürmuhl and Stevens indicate that at frequencies above 2000 cycles there is a small increase in pitch as the level is raised. An observer's judgment of the pitch of a sound is thought to be related to the position of the stimulated region on the basilar membrane. In general, a sensation of lower pitch occurs when the stimulated region shifts towards the helicotrema.

Table 6. Judgments of Half Pitch

(Stevens and Volkmann)

Standard frequency	150	250	500	1,000	2,000	3,000	5,000	10,000
Frequency for half pitch	85	111	206	373	633	1,009	1,437	2,064

Experiments have been made to determine how much the frequency of a standard tone must be shifted to result in the sensation that the pitch has decreased by a factor of one-half. The results are shown in Table 6, where the first row is the frequency of the standard tone. The second row is the mean frequency of tones which were selected by twelve

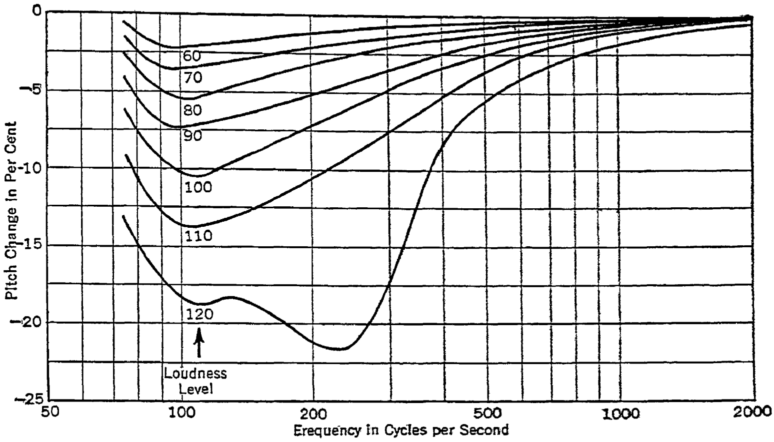


FIG. 20. Pitch Change of Single-frequency Tones at High Levels. (Snow.)

observers to be one-half the pitch of the standard. The loudness level of all tones was 40 db. Table 6 shows that a tone of 373 cycles will appear to be one-half as high in pitch as 1000 cycles. By use of these and other data on estimations of pitch intervals, Stevens, Volkmann, and Newman have devised a numerical pitch scale having the property that tones which appear to be 50 per cent lower in pitch will also be related in the same manner on the pitch scale. The unit of pitch is called a "mel," and the relationship between mels and frequency for pure tones at a loudness level of 40 db is shown in Fig. 21.

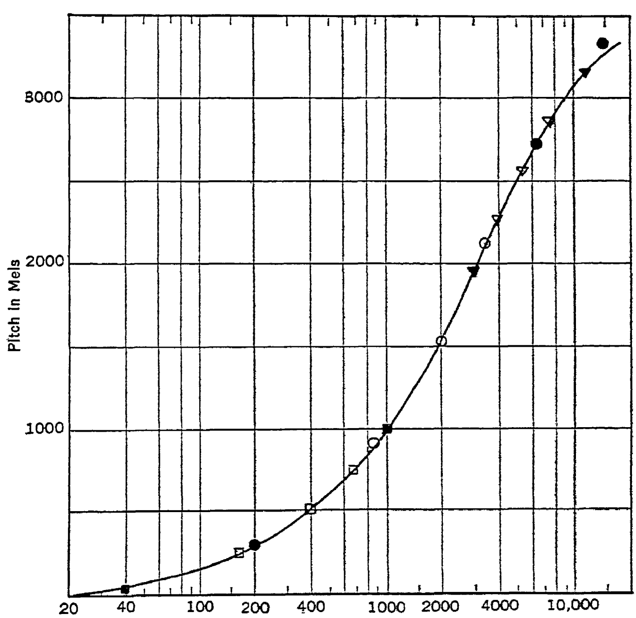


FIG. 21. Relation of the Pitch in Mels to the Frequency of Tones at a Loudness Level of 40 Db. (Stevens and Volkmann.)

7. LOCALIZATION OF SOUNDS

The ability to localize the direction and to form a judgment of the distance away of a source is a matter of common experience, but very few quantitative data on the subject are available. The localization of direction (angular localization) appears to depend upon the detection of phase differences at the two ears, loudness differences, differences in quality, and differences in arrival times at the two ears. Loudness differences, and quality differences (in complex sounds), and to some extent phase differences, arise from diffraction effects of the head. At low frequencies the loudness difference at the two ears is small, but localization by phase difference is effective. An angular accuracy of about $\pm 10°$ is obtained for sounds directly in front or in back of the observer. Much less precision is obtained when the source is at one side. Loudness differences are most effective in the frequency range above 3000 cycles and quality differences in the range above 1000 cycles. In general, complex sounds having prominent high-frequency components, hence large loudness and quality differences between the ears, are localized with greatest accuracy. Under familiar acoustic conditions, such sounds may be localized frequently by the use of one ear only. Presumably this is accomplished by recognizing the characteristic distortion introduced by the head. Steinberg and Snow report that the apparent distance of the sound source, i.e., depth localization, depends upon the loudness of a sound. When an observer is listening in a room, the ratio of the direct sound intensity (that reaching the ears without reflection) to the reflected sound intensity also enters into depth localization.

BIBLIOGRAPHY

American Standards Association, American Standard Acoustical Terminology, *J. Acous. Soc. Am.*, Vol. 14, 84 (July 1942).

Beasley, W. C., Characteristics and Distribution of Impaired Hearing in the Population of the U. S., *J. Acous. Soc. Am.*, Vol. 12, 114 (July 1940).

Bekesy, G., Clicks and the Theory of Hearing, *Physik. Zeitschrift*, August 1934, p. 577.

Dimmick and Olson, Intensive Difference Limen in Audition, *J. Acous. Soc. Am.*, Vol. 12, 517 (April 1941).

Fletcher, H., Loudness, Masking and Their Relation to the Hearing Process and the Problem of Noise Measurement, *J. Acous. Soc. Am.*, Vol. 9, 275 (April 1938).

Fletcher, H., Loudness, Pitch, and Timbre of Musical Tones, *J. Acous. Soc. Am.*, Vol. 6, 59 (October 1934).

Fletcher, H., A Space-time Pattern Theory of Hearing, *J. Acous. Soc. Am.*, April 1930, p. 311.

Fletcher, H., Auditory Patterns, *Rev. Modern Phys.*, Vol. 12, 47 (January 1940).

Fletcher, H., *Speech and Hearing*. Van Nostrand (1929).

Fletcher, H., The Mechanism of Hearing, *Proc. Natl. Acad. Sci.*, Vol. 24, 265 (July 1938).

Fletcher and Munson, Loudness, Its Definition, Measurement, and Calculation, *J. Acous. Soc. Am.*, Vol. 5, 82 (October 1933).

Fletcher and Munson, Relation between Loudness and Masking, *J. Acous. Soc. Am.*, Vol. 9, 1 (July 1937).

Galambos and Davis, The Response of Single Auditory-nerve Fibers to Acoustic Stimulation, *J. Neurophysiol.*, Vol. 6, 39 (1943).

Inglis, Gray, and Jenkins, A Voice and Ear for Telephone Measurements, *Bell Sys. Tech. J.*, April 1932, p. 293.

Knudsen, V. O., Sensibility of the Ear to Small Differences of Intensity and Frequency, *Phys. Rev.*, January 1923, p. 84.

Rawdon-Smith and Sturdy, The Effect of Adaptation on the Differential Threshold for Sound Intensity, *Brit. J. Psych.*, Vol. 30, 124 (1939).

Riesz, R. R., Differential Intensity Sensitivity of the Ear for Pure Tones, *Phys. Rev.*, May 1928, p. 867.

Shower and Biddulph, Differential Pitch Sensitivity of the Ear, *J. Acous. Soc. Am.*, October 1931, p. 275.

Sivian and White, On Minimum Audible Sound Fields, *J. Acous. Soc. Am.*, April 1933, p. 288.

Snow, W. B., Change of Pitch with Loudness at Low Frequencies, *J. Acous. Soc. Am.*, Vol. 8, 14 (July 1936).

Steinberg, J. C., Positions of Stimulation in the Cochlea by Pure Tones, *J. Acous. Soc. Am.*, Vol. 8, 176 (January 1937).

Steinberg, Montgomery, and Gardner, Results of the World's Fair Hearing Tests, *J. Acous. Soc. Am.*, Vol. 12, 291 (October 1940).

Steinberg and Snow, Auditory Perspective—Physical Factors, *Elec. Eng.*, Vol. 53, 12 (January 1934).

Stevens, S. S., The Attributes of Tones, *Proc. Natl. Acad. Sci.*, July 1934, p. 457.

Stevens, S. S., and H. Davis, *Hearing, Its Psychology and Physiology*. John Wiley (1938).

Stevens and Davis, Psychophysiological Acoustics—Pitch and Loudness, *J. Acous. Soc. Am.*, Vol. 8, 1 (July 1936).

Stevens and Volkmann, The Relation of Pitch to Frequency: A Revised Scale, *Am. J. Psych.*, Vol. 53, 329 (July 1940).

Stevens, Volkmann, and Newman, Scale for the Measurement of Pitch, *J. Acous. Soc. Am.*, Vol. 8, 185 (January 1937).

Stewart, G. W., Localization of Pure Tones, *Phys. Rev.*, May, 1920, p. 425.

Trimmer and Firestone, An Investigation of Subjective Tones, *J. Acous. Soc. Am.*, Vol. 9, 24 (July 1937).

Tröger, J., The Reception of Sound by the Outer Ear, *Physik. Zeitschrift*, January 1930, p. 26.

Wallach, Hans, On Sound Localization, *J. Acous. Soc. Am.*, Vol. 10, 270 (April 1939).

Wegel and Lane, Auditory Masking and Dynamics of the Inner Ear, *Phys. Rev.*, February 1924, p. 266.
Wever, E. G., The Electrical Responses of the Ear, *Psych. Bull.*, March 1939, p. 36.
Wrighton, Sir Thomas, *Analytical Mechanism of the Internal Ear.* Macmillan, London (1918).
Zürmuhl, G., Variation of Pitch Sensation with Loudness, *Zeit. für Sinnesphysiologie*, August 1940, p. 40.

SPEECH AND MUSIC

By John C. Steinberg and W. A. Munson

8. DESCRIPTION OF SPEECH ORGANS

The speech organs consist of the lungs and respiratory muscles, the trachea or windpipe, the larynx, and the cavities of the throat, mouth, and nose. In speaking, a flow of air is produced by the lungs which is modified in passing through these passages to form speech sounds. The larynx is a cavity of irregular shape formed of cartilage located at the upper end of the trachea, a tube some 12 cm long and 2 cm in diameter leading from the lungs. A pair of muscular ledges, called the vocal cords, form a slit in the larynx through which the air must pass. When the vocal cords are set in vibration, the air flow is periodically interrupted and the sound is said to be voiced. Measurements reported by Riesz, on persons whose larynx had been removed by a surgical operation, indicate that the excess pressure in the lungs when producing sustained parts of speech, such as vowels, is of the order of 4 mm of mercury or 0.005 atmosphere. The rate of air flow is about 150 cu cm per sec. The normal capacity of the lungs is about 2500 cu cm, and the average expiration in breathing is about 500 cu cm.

By means of high-speed motion-picture photography, views of the vocal-cord movements in slow motion have been obtained at Bell Telephone Laboratories. Measurements on successive exposures of the vibrating cords indicated a displacement amplitude that was sawtooth in form. The cords tended to snap apart and close slowly and firmly together. The wave form contained a fundamental and higher harmonics which diminished in amplitude inversely as the square of the harmonic number. The length of the cords and maximum amplitude of opening varied with pitch. For example, in the subjects studied, the cords were about 1.2 cm in length and their widest opening was about 0.4 cm when vibrating at 120 cycles. At 300 cycles, their length was about 2.0 cm and the widest opening about 0.2 cm. Changes in the acoustic load by closing the mouth opening with a glass window or filling the vocal cavities with helium did not appear to affect the vibration wave form of the vocal cords markedly.

The opening between the cords is called the glottis, which, of course, varies in size as the cords vibrate. There appears to be little direct information on the acoustic wave form produced at the glottis by the vocal-cord vibrations. Indirect evidence based on harmonic analysis of vowel sounds reported by Steinberg and also Lewis indicates a sawtooth wave form in which the amplitudes of the harmonics diminish inversely as the 1.5 power of the harmonic number.

9. PRODUCTION OF SPEECH

MECHANISM OF SPEECH. As has been discussed by Dudley, speech may be regarded as a phenomenon of modulation in which the breath stream, a unidirectional flow of air produced by the lungs, is given intelligence-carrying variations. The breath stream is made audible by two kinds of modulation called *vocal-cord* and *frictional* modulation. The first is produced by the periodic interruption of the breath stream by the vibrations of the vocal cords. The second is produced by the turbulent flow of the breath stream through constrictions formed in the vocal tract. Both produce variations at relatively high rates, i.e., in excess of 70 cycles. The vocal-cord modulation produces fundamental and harmonic overtones such as characterize a vocalized or voiced sound. Frictional modulation produces a wide range of inharmonically related overtones such as characterize a hiss.

The audible components thus produced are in turn modulated by relatively slow variations, i.e., at rates definable by frequency components in the range below some 40 or 50 cycles. One of these is called *start-stop* modulation, which is accomplished by the muscles of respiration, the vocal cords, tongue, and lips. Another, called *cavity* modulation, is produced by changes in size and shape of the vocal passages extending from the glottis to the mouth and nose openings. These passages selectively transmit and radiate the audible frequency components produced in the breath stream by vocal-cord and frictional modulation. This results in a reinforcement, relatively, of certain frequency components or regions which are called vocal resonances.

Several other types of low-frequency modulation occur such as inflection, vibrato, and stress, but the four noted above are the ones of chief importance to the intelligibility of American speech sounds. Start-stop modulation is one of the principal characteristics of the plosive or stop consonant group, and frictional modulation is one of the chief characteristics of the fricative consonant group. In the vowel and vowel-like group of sounds, cavity and vocal-cord modulation are among the conspicuous characteristics.

CHARACTERISTICS OF SPEECH SOUNDS. As a result of cavity modulation, all sounds tend to show characteristic frequency regions of reinforcement in a greater or less degree. These vocal resonances are more conspicuous in the vowel and vowel-like sounds. In general, there is a resonance in the range from 300 to 1000 cycles, one in the range from 1000 to 2500 cycles, and one or more in the range from 2500 to 4000 cycles. The two resonances below 2500 cycles are the ones that vary most in frequency position from sound to sound. There has been a tendency to associate the low resonance with the pharynx or throat cavity and the next resonance (1000–2500 cycles) with the mouth cavity. Damping constants reported for the cavities range from 1000 to 4000 db per sec. With a vocal-cord vibration rate of 125 cycles, the cavities would be excited periodically at intervals of 0.008 sec, and the amplitude of the sound wave would be expected to decay to something between 0.4 and 0.03 of its initial value during the interval.

Table 1. Approximate Characteristics of Common Speech Sounds

Sound	f_1	L_1	f_2	L_2	f_3	L_3	f_4	L_4	n	
Pure Vowels										
ē (eve)	350	−5	2400	−17	3200	−13	3700	−23	6.44	
i (it)	450	−4	2150	......	2950	......	3600	......	10.27	
e (bet)	550	−3	1950	−9	2700	−13	3700	−19	6.60	
ă (at)	800	−1	1800	......	2800	......	3900	......	6.89	
ah (father)	800	0	1250	−8	2950	−22	3800	−30	6.52	
a (all)	550	0	850	......	2900	......	......	......	4.15	
ō (obey)	450	−2	800	−9	2600	−24	3200	−24	4.74	
u (foot)	400	−2	1050	−12	2300	−24	3200	−34	2.96	
ū (boot)	350	−4	950	−19	2250	−32	3200	−34	6.26	
Vowel-like Sounds									n_i	n_f
l (let)	450	−8	1000	−21	2550	−20	2950	−24	4.31	8.40
r (run)	500	−5	1350	−12	1850	−16	3500	−29	2.78	13.05
m (me)	Fundamental	−11	1250	−21	2250	−23	2750	−30	5.89	5.48
n (no)	"	−13	1450	−26	2300	−28	2750	−33	4.99	12.52
ng (sing)	"	−10	2350	......	2750	......				3.57
Fricative Consonants										
v (voice)	Fundamental	−18	1150	−37	2500	......	3650	......	1.2 (3.9)	4.2 (1.4)
th (that)	"	−18	1450	−27	2550	−45	......	......	6.7 (2.0)	1.2 (0.04)
z (zoo)	"	−16	2000	−33	2700	−42	......	......	0.3 (5.4)	6.0 (3.1)
zh (pleasure)	"	−15	2150	−24	2650	−39	......	......	0.02 (1.7)	0.01 (0.3)
Stop Consonants										
b (be)	Fundamental	−20	800	......	1350	......	......	......	4.6 (2.5)	0.4 (1.2)
d (day)	"	−20	1700	......	2450	......	......	......	6.2 (7.8)	4.4 (14)
g (get)	"	−17	← Variable					→	4.3 (5.5)	0.4 (2.8)

The steady-state resonance characteristics of the sounds can be expressed by the peak frequencies f_i of the vocal resonances, the levels L_i of the peak frequencies, and the damping constants Δ_i expressed in decibels per second which indicate the sharpness of the resonances. This information, together with the data on the acoustic wave form at the glottis given in article 8, permits an approximate construction of the spectrums of the sounds. The available data on these characteristics are summarized in Table 1. The columns designated f_1 to f_4 indicate the peak frequencies of vocal resonance. They were obtained from data reported by Kopp and Green, and represent values for one voice. Since the values vary somewhat from voice to voice, values for a single voice were chosen to permit a relative comparison of the different sounds. When "fundamental" appears in the column f_1, it means that the resonance was too close to the fundamental frequency to be resolved by the analyzing means. The vocal resonances shown for the voiced consonants, *v*, *th* (that), *z*, *zh*, *b*, *d*, and *g*, apply also to their unvoiced cognates, *f*, *th* (thin), *s*, *sh*, *p*, *t*, and *k*. For *g* (get) and also *k* (key), the frequencies of the vocal resonances vary depending upon the sound with which they are combined. The column L_1 shows the relative levels of the first resonance peaks as derived from Table X, Fletcher, *Speech and Hearing*.

The remaining columns, L_2 to L_4, show levels of the respective resonant peaks based on data from several sources. All levels are expressed in decibels below the level of the first resonant peak in *ah* (father). The damping constants Δ_i are not shown but vary from about 1200 db per sec for the first resonance to about 4000 db per sec for the fourth resonance. The band width of the resonance at points on the resonance curve 3 db below the peak response is given approximately by $\Delta_i/27.3$. The last column, designated n, shows the relative occurrence of the sounds in telephone conversation from a report by French, Carter, and Koenig. The figures indicate the percentage occurrence of the vowels shown among all vowels occurring in 95,522 vowels. The occurrence figures are indicated separately for initial and final consonants from 64,043 initial and 65,544 final consonants. The figures in parentheses indicate the occurrence of the unvoiced cognates of the voiced consonants shown. The figures were obtained from the words, syllables, and sounds occurring in telephone conversations after the exclusion of articles, names, titles, exclamations, letters, and numbers.

An interesting feature of the report is the extent to which a few common words or sounds, by being used over and over, from a large part of ordinary speech. For example, eight different words and four different sounds account, respectively, for 25 per cent of the total words and sounds used. Thirty words and ten sounds account, respectively, for 50 per cent of the total used.

ARTIFICIAL LARYNX. The artificial larynx is based on the principle that the vocal-cord tone need not necessarily arise in the larynx in order that cavity modulation occur to form speech. It may be introduced into the vocal cavities through the mouth opening. A means for accomplishing this has been described by Riesz and is used by persons who have had their larynx removed by surgery. In this case, the windpipe is terminated by a small opening in the neck, through which the patient breathes. In using the artificial larynx, a flexible rubber tube is fitted over the neck opening and leads to a reed which vibrates with the passage of air. The vibrated air stream is led to the mouth opening by a flexible tube, and, with practice, the patient learns to modulate the sound and form speech by the ordinary articulatory movements. Firestone has described and demonstrated a form of artificial larynx for introducing various types of sounds through the mouth opening for vocal modulation. Mr. G. M. Wright has developed an artificial larynx in which sound is introduced into the vocal tract by means of vibrators attached outside the throat. The device has been used to produce unusual vocalized sound effects in radio work.

ARTIFICIAL VOICE. An artificial voice described by Inglis, Gray, and Jenkins, designed for the measurement of microphones, consists of a small moving-coil loudspeaker having an opening somewhat larger than that of the human mouth. Its design provides for an undistorted acoustic power output comparable with powers produced in speaking. Its sound field approximates that of the human mouth to the extent that the loss in power delivered by a microphone with increasing distance between the mouth and the microphone is about the same for the artificial as for the human voice.

THE VOCODER. A means is described by Dudley for continuously analyzing speech and utilizing the results of the analysis to synthesize or remake the speech. It is based on the principle that the intelligibility of speech is carried by the relatively low-frequency components produced by cavity modulation and that the "buzz tone" (vocal-cord modulation) and the "hiss tone" (frictional modulation) simply act as carrier waves which are modulated by the vocal cavities. Frequency range reduction is achieved by transmitting only the low-frequency modulations and employing them to modulate locally generated buzz and hiss tones. In one variation called the voder, speech is produced artificially by using a keyboard manipulated with the fingers for generating the carrier waves and the low-frequency modulations.

VISIBLE SPEECH. A development described by Potter, in which speech is continuously analyzed and the results of the analysis portrayed to the eye in the form of visible patterns that one can learn to read, holds possibilities in the fields of visual hearing and visual telephony for the deaf, phonetic printing and retranslation into sound, the selective operation of automatic devices by voice sounds, and in the specialized fields of phonetics, linguistics, foreign language, music, etc To obtain readable patterns of speech sounds, the portrayal emphasizes the modulations that are important to intelligibility as illustrated in Fig. 1. Frequency is shown vertically, time horizontally, and intensity by shades of gray. The stop gap (start-stop modulation), the plosive release (frictional modulation), and the voice bar (vocal-cord) modulation are indicated for the voiced stop consonant *b*. Likewise the vocal resonance bars (cavity modulation) are indicated for the vowel *ē*. The *u* shows only two resonance bars which differ in position from those of the *ē*. The transition from *ē* to *u* is shown also. The fricative fill for *f* is an example of frictional modulation. The patterns for voiced sounds show regular vertical striations, the space between the striations indicating the pitch. Note the drop in pitch toward the end of "five." The unvoiced sounds (frictional modulation) show irregularly spaced vertical striations indicating a lack of pitch. The patterns are read from the characteristic modulations of the sounds and the transitions between them.

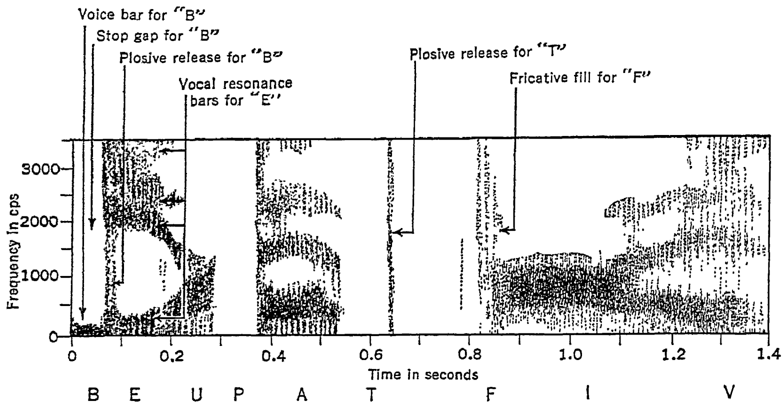


FIG. 1. Visible Patterns of the Words "Be up at Five"

10. SPEECH POWER

Since the wave forms of speech are complex, different types of speech power have been defined as follows:

Instantaneous. The rate at which sound energy is being radiated by the speaker at any given instant.

Average. The average speech power for any given time interval is the average value of the instantaneous speech power over that interval.

Peak. The maximum value of the instantaneous speech powers occurring in a given time interval.

Phonetic. The phonetic speech power is the maximum value of the average speech power in 0.01-sec intervals of a vowel or consonant sound. It is the maximum value of the envelope that results from plotting average power for 0.01-sec intervals against time, as the sound grows, remains steady, and decays.

The quantity usually measured is the speech pressure at some convenient distance and direction from the mouth. Dunn and White have reported the results of measurements of various types of speech pressure and comparisons with earlier measurements, and Dunn and Farnsworth have reported on the directional characteristics of the mouth as a radiator. The results are summarized in Table 2. They comprise statistical measurements of speech pressure vs. frequency range for conversational speech made with an arrangement for introducing any one of 14 band-pass filters into a speech circuit. The frequency range below 500 cycles was covered in one-octave steps. Above 500 cycles, each filter passed a range of about $^1/_2$ octave. Provisions were made for measuring instantaneous, peak, and rms pressures in alternate time intervals of $^1/_8$ sec and also, for certain cases, in time intervals of 15 sec. Some 600 intervals of $^1/_8$ sec were usually measured for a given condition representing an integrated time of 75 sec.

From the results it is possible to calculate the various types of speech power. This is done by converting pressure to power per square centimeter (intensity) by the relation $I = p^2/415$, where I = intensity in microwatts per sq cm and p = pressure in dynes per sq cm at a given distance and direction from the mouth. The intensities in other directions are obtained by weighting in accordance with the directional characteristics of the mouth, and P, the total power radiated, is then obtained by integrating over a spherical surface having a radius equal to the distance. The results of such calculations are given in the first row of Table 2 in the form of ratios expressed in decibels. The ratios depend upon the distance and direction from the mouth and the frequency range in which p is measured. The reference position is designated as 30, 0°, 0°, signifying 30 cm from the mouth at zero azimuth and altitude angles, i.e., directly in front of the mouth. The values when added to p expressed in decibels from 1 dyne per sq cm give P expressed in decibels from 1 microwatt. The values are based on distribution measurements for one voice.

Table 2. Summary of Data on Speech Powers

Quantity		Whole Spectrum	62.5–125~	125–250~	250–500~	500–700~	700–1000~	1000–1400~	1400–2000~	2000–2800~	2800–4000~	4000–5600~	5600–8000~
P/p^2 in db, p in dynes/cm^2 at 30, 0°, 0°. P in microwatts		13.3	14.7	13.5	12.9	14.5	13.5	10.8	10.0	10.5	10.3	8.7	10.3
Long-interval rms p at 30, 0°, 0°, db from 1 dyne/cm^2	Men	2.0	−14	−7	−2	−5	−12	−12	−13	−18	−21	−28	−29
	Women	−0.8	....	−8	−5	−10	−13	−16	−16	−22	−25	−25	−26
p in different directions expressed in db from p at 30, 0°, 0°.	30, 0°, −45° (down)	1.0	1.6	1.7	1.1	2.9	1.6	−1.1	−1.8	−1.1	−2.9	−0.5	2.1
	30, 0°, +45° (up)	−0.8	−0.6	−1.0	−1.9	−1.3	−0.5	1.1	−1.4	0.2	−0.7	−4.4	−4.3
	30, ±45°, 0° {right, left}	−0.1	−0.1	0.6	0.1	0.0	0.0	−1.2	−0.8	−2.5	0.6	−2.9	0.5
	30, 180°, 0° (back)	−4.9	−0.6	−4.6	−5.3	−3.3	−3.8	−5.1	−12.4	−14.8	−17.2	−18.6	−21.3
Peak pressure in 1/8-sec intervals expressed in db from long-interval rms p at 30, 0°, 0°. Peaks exceeded in 1 and 10% of intervals	1%	20	16	14.5	17	19	22	23	23	25	27	28	27
	10%	16	11	10	13	14	15	15	18	20	21	20	16
Distribution of rms p in 1/8-sec intervals expressed in db from long-interval rms p. % of intervals in which 1/8-sec rms p exceed long-interval rms p by the indicated amounts	% of intervals			1	5	10	20	30	40	50	60	70	80
	db from long-interval rms p whole spectrum			9.0	5.5	4.0	2.0	0.5	−1.0	−3.0	−6.0	−9.0	−13.0
	db from long-interval rms p 700–1000~ band			12.0	7.0	4.0	0	−3.0	−6.0	−9.0	−12.0	−15.0	−18
Distribution of instantaneous and 1/8-sec peak p. % of time that instantaneous and peak p exceeds long-interval rms p by indicated amounts.	% of time			1	5	10	20	30	40	50	60	70	80
	db from long-interval rms p whole spectrum	Inst.		20	6	2	−2	−5.5	−9	−13.5	−18.5	−25	−32
		Peak		20	18	16	14	11	9	6	4	0	−6
Distribution of long-interval rms p among speakers. % of speakers for which long-interval rms p exceeds average long-interval rms p by indicated amounts	% of speakers			1	5	10	20	30	40	50	60	70	80
	db from average long-interval rms p			9	5.5	4.0	1.5	0	−1.0	−2.5	−4.0	−6.0	−8.0

The second and third rows give the long interval rms pressure p at 30, 0°, 0° in decibels from 1 dyne per sq cm. The rows designated "men" and "women" are averages for 6 male and 5 female speakers, respectively. By long interval is meant a time average over some 600 $^1/_8$-sec intervals or 75 sec. Corresponding values in other bands may be obtained by taking p^2 as proportional to band width in cycles per second. Rows 4, 5, 6, and 7 show pressures p at different directions relative to the reference position. The mouth radiates maximally in a direction about 45° down from the horizontal. Rows 8 and 9 show peak pressures that are exceeded in 1 and 10 per cent of the $^1/_8$-sec intervals relative to the long-interval values. One per cent peaks more than 20 db above long-interval pressures occur frequently. Corresponding values for other bands may be obtained approximately for peaks occurring less than 10 per cent of the time, by proceeding as though they were rms pressures. The peak-factor data are essentially the same for both men's and women's speech.

Rows 10, 11, and 12 show the distribution of $^1/_8$-sec rms pressures relative to the long-interval values given in rows 2 and 3. Distributions for the three low-frequency bands are very similar to the whole spectrum distribution, row 11. Distributions for the remaining bands are very similar to that for the 700 to 1000 cycle band given in row 12. These distributions are about the same for both men's and women's speech.

Rows 13, 14, and 15 give distributions of instantaneous and peak pressures relative to the long-interval rms pressure. The distribution for instantaneous pressures is based on data for a single male voice.

The last two rows give data on the distribution of long-interval rms pressures among various speakers reported by Fletcher.

The results in Table 2 indicate a long-interval total speech power, averaged for 6 men, of 34 microwatts. The corresponding figure averaged for 5 women is 18 microwatts. These values are somewhat larger than that of 10 microwatts given formerly. Some 2 db of this difference is due to the method of converting measured pressures to radiated powers. These values obtain for continuous connected speech. If the silent intervals (about $^1/_5$ to $^1/_3$ of the total time) are excluded, the average is increased about 25 to 50 per cent. The above values hold for about 30 per cent of the speakers. One per cent of the speakers may radiate powers in excess of 272 microwatts. If one shouts as loudly as possible, the total power may reach 3400 microwatts, an increase of 20 db. For a very faint but intelligible whisper, the total power may fall to 0.0034 microwatt, a drop of 40 db.

In 1 per cent of $^1/_8$-sec intervals, the peak power may exceed the long-interval total power in connected speech by 20 db. Thus total peak powers of the order of 3400 microwatts may be reached by average male voices.

The *ah* (father) is about the loudest sound in speech. Fletcher reports a total phonetic power of 41 microwatts for this sound. Corresponding values for other sounds may be obtained from the column designated $L1$ of Table 1. These obtain for discrete words or syllables spoken at conversational level. Peak powers of the order of 1600 microwatts for the *ah* sound were obtained under these conditions. Wolf, Stanley, and Sette report total phonetic powers of 1 watt when the vowel *ah* is sung by professional singers. To be consistent with the data reported in Table 2, the total phonetic powers given above should be increased by 2 db.

Sivian reports on changes in the frequency distribution of speech power as a speaker changes from a low (confidential) talking level to a normal (conversational) level and to a high (declamatory) level. The total change in power from low to high was about 24 db. In general, power was transferred from the frequency range below 500 cycles to the range between 500 and 4000 cycles as the talking level was raised. To be representative of declamatory speech, the values in the bands below 500 cycles given in row 2, Table 2, should be decreased about 4 db and values between 500 and 4000 cycles should be increased about 3 db.

11. POWERS PRODUCED BY MUSICAL INSTRUMENTS

In music, as in speech, various aspects of power—peak, average, etc.—must be dealt with. A comprehensive report on the powers produced by the various musical instruments and by orchestras has been given by Sivian, Dunn, and White. The report also describes in some detail the band-pass filter apparatus discussed in article 10 on speech power. Two types of measurements were made on the waves, the average pressure in 15-sec intervals and the peak pressures in $^1/_8$-sec intervals. The pressures are the field pressures at the position in the field where the microphone of the measuring circuit was placed. For the measurements on individual instruments, the position of the microphone with respect to the source varied with instrument. For the orchestra, the microphone

Table 3. Peak Powers in Music

Instrument	Microphone Position and Assumption in Converting to Total Sound Power	Field Pressure dynes/cm²		Total Peak Power Watts	Percentage of Intervals	Band Containing Maximum Peaks	Field Pressure dynes/cm²		Total Peak Power Watts	Percentage of Intervals
		Average	Peak				Average	Peak		
Bass drum 36″ × 15″	3 ft in front, on axis. Radiation confined to a cylinder having drum diameter.	99.0	1260.0	24.6	6.0	250–500	35.0	792.0	9.8	1.0
Bass drum 30″ × 12″		35.0	980.0	13.4	1.0	125–250	12.0	392.0	1.7	1.0
Snare drum	4 ft in front, 90° off axis. Peak pressure increased 8.5 db for 1-ft distance. Radiation confined to hemisphere.	14.6	365.0	11.9	2.5	250–500	8.8	193.0	3.7	1.0
15″ cymbals	3-ft distance. Peak pressure increased 7.2 db for 1 ft. Radiation confined to hemisphere.	18.0	360.0	9.5	7.5	8,000–11,300	2.5	128.0	0.95	1.0
Triangle	3-ft distance. Conversion as for cymbals.	2.3	25.8	0.05	1.0	5600–8000	0.9	11.5	0.017	6.0
Bass viol	3-ft distance. Radiation confined to hemisphere.	4.2	37.8	0.156	2.0	62–125	3.1	26.4	0.078	3.0
						125–250	0.8	26.4	0.078	2.0
Bass saxophone	3-ft distance. Radiation confined to hemisphere.	4.1	58.2	0.288	25.0	250–500	3.3	51.3	0.228	4.0
BB♭ tuba	3-ft distance. Conversion made from measurements with a complex sound source attached to a horn of similar size.	5.4	43.2	0.206	17.0	250–500	3.5	27.0	0.082	18.0
Trombone	3-ft distance. Conversion as for tuba.	6.5	228.0	6.4	5.0	500–700	2.9	22.8	0.064	1.0
						2000–2800	0.04	20.5	0.051	4.0
Trumpet	3-ft distance. Conversion as for trombone.	8.6	54.2	0.314	18.0	250–500	2.4	19.3	0.047	1.5
						500–700	1.5	19.3	0.047	4.5
French horn	As for trumpet.	3.8	27.0	0.053	6.0	250–500	3.8	27.0	0.053	1.0
								13.5	0.013	18 0
Clarinet	As for trumpet.	3.3	26.4	0.050	5.5	250–500	0.7	8.2	0.0055	2.5
Flute	As for trumpet.	1.6	25.6	0.055	1.0	700–1000	0.6	7.2	0.0045	4.0
			12.8	0.014	1.5					
			6.4	0.0035	38.0	1400–2000	0.2	7.2	0.0045	2.0
Piccolo	As for trumpet.	2.2	30.8	0.084	0.5	2000–2800	0.4	15.4	0.021	3.0
			15.4	0.021	10.0					
Piano	10-ft distance. Room 29′ × 29′ × 13′. Reverberation time 1 sec, 60 to 4000~, average of 3 methods (see text).	2.6	23.4	0.267	16.0	250–500	2.1	23.4	0.267	7.0
15-piece orchestra	6 ft from nearest instruments, in same room as piano. Average of 2 methods (see text).	7.9	126.0	9.0	1.5	250–500	2.8	27.6	0.45	1.5
			63.2	2.2	16.0	2000–2800	1.3	22.1	0.32	12.0
75-piece orchestra	15 ft from nearest instrument in theater (see text).	4.6	129.0	66.5	1.0	250–500	1.6	5.1	6.7	1 0
						8,000–11,300	0.2	4.6	5.3	1.0
Pipe organ	Effective distance 15 ft. Radiation assumed uniform over 1/4 sphere.	20.0	90.0	12.6	36.0	20–62.5	19.0	80.0	10.0	1.0
								40.0	2.5	22.0

was placed near the conductor's stand. By making assumptions as to the radiating properties of the instruments, and of the orchestra, their total power outputs could be estimated.

Table 3 gives the average and peak pressures at the positions of measurement as indicated. The columns "total peak power" give the total power radiated by the instruments, computed on the basis of assumptions indicated in the table. The left half of the table applies to the whole spectrum; the right half to the bands containing the maximum peaks.

An average of three methods was used for determining the total power radiated by the piano. In the first method, the diffuse pressure in a room of known reverberation was measured and the power calculated. In the second method, from auxiliary measurements, it was shown that the measured diffuse pressure was the same as if the total power output were distributed over a sphere of radius 5.37 ft. In the third method, the diffuse pressure was compared with the pressure measured in the opening between the raised top (grand piano) and the sounding board. The radiated power was assumed to be uniform over the area of the opening, 3 ft by 6 ft. The three methods gave peak powers of 0.166, 0.437, and 0.198 watt, respectively.

The measurements on the 15-piece orchestra were made in the same room as the piano measurements, and methods 1 and 2 were used to obtain the total power. The sounds from the 75-piece orchestra were picked up in a theater, and suitable acoustic data for using the above methods were not available. It was assumed that the radiation was uniform over a hemisphere of 15-ft radius. The measurements on the pipe organ were made in the same theater.

Because of the uncertain assumptions in calculating total powers, the measured values of pressure are also given in Table 3. The peak pressures were measured within a range of ±3 db, and the percentage of $^1/_8$-sec intervals that the power fell within this range is indicated. The peak pressures are useful in determining the amplitudes that pick-up instruments and amplifiers must handle, and the total peak powers for determining the requirements that must be met by power amplifiers and loudspeakers when reproduction of the original level is desired. The most powerful single instrument is a bass drum, which radiates a power of about 25 watts, primarily in the low-frequency ranges. A large orchestra is capable of radiating power of about 60 to 70 watts. This is about 70,000 times the peak power of speech.

Figure 2 (from Fletcher) shows the frequency distribution of the maximum and most probable peak powers for a 75-piece orchestra. The curves are based on average measurements of four selections which gave whole-spectrum peak powers from 8 to 66 watts, and

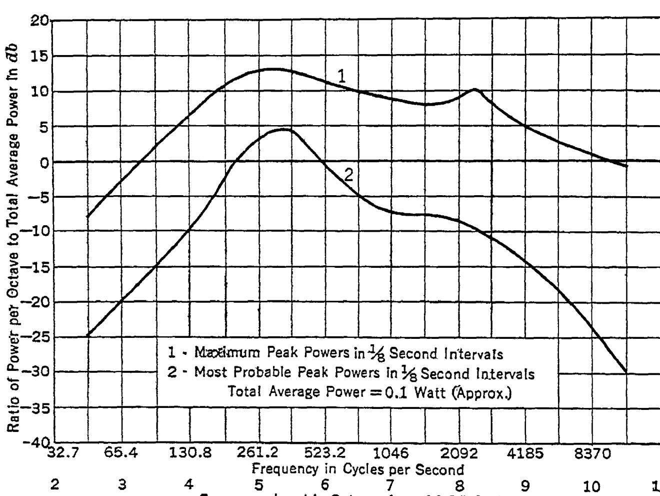


Fig. 2. Maximum and Most Probable Peak Powers

average powers from 0.08 to 0.13 watt. The zero line corresponds to an average power of about $^1/_{10}$ watt. The ordinate is the ratio expressed in decibels of the peak power per octave to the whole-spectrum average power.

A violin player, when asked to play at the lowest level that could be used with an audience, produced a whole-spectrum average pressure of 0.52 dyne per sq cm for a 3-ft distance. The highest peak observed for the bass drum was 1260 dynes per sq cm. Thus a range of at least 68 db is indicated between the highest and lowest amplitudes in music. The estimated range probably errs in the direction of being too small rather than too large.

12. TESTS OF SPEECH AND MUSIC TRANSMISSION

The performance of a system for the transmission and reproduction of speech or music can be expressed in terms of objective measurements involving a determination of its characteristics for steady-state sinusoidal waves of different frequency, or in terms of subjective measurements involving the satisfactoriness of the system from the viewpoint of the person receiving the signals. This section is concerned with the subjective type of measurement which involves, for speech, the recognizability and the naturalness of the reproduced sounds, and for music, the emotional and esthetic properties of the reproduction. Measurements of this kind may be classed as laboratory and field measurements. The former are particularly adapted to fundamental studies of the effects of various physical factors on the reproduction; the latter, to evaluating performance under conditions of actual use.

LABORATORY TESTS. The articulation test has come into wide use as a laboratory method of measuring the recognizability of received speech sounds. One form of the test which was used in obtaining the data given in the next article has been described by Fletcher and Steinberg. In this form, a speaker utters the various speech sounds and an observer writes down the sounds which he hears. The observed sounds are compared with the called sounds, and the percentage of the called sounds that were correctly recognized is obtained. This percentage is the articulation. To call the sounds, they are combined in a random manner into syllables of the consonant-vowel-consonant type which have no meaning in English. The syllables are spoken as parts of introductory sentences, such as, "Please record the syllable ——." The sounds used are those shown in Table 1 (p. 12-20), and they occur with uniform frequency in the testing lists. The percentage of the total number of spoken syllables which are correctly observed is called the "syllable articulation." The "sound articulation" is the percentage of the total number of spoken sounds which are correctly observed. When attention is directed toward a specific fundamental sound, e.g., *b*, the term "individual sound articulation" is used. Similarly, "vowel articulation" is the percentage of the total number of spoken vowel sounds which are correctly observed.

To obtain reproducible and representative results it is important that the testing crew have normal voices and hearing and that they be thoroughly familiar with the method. The best number of voices and observers and the amount of testing material required depend on the discrimination which is desired. For most of the data reported in the next article, each of 8 callers reads a list of 66 syllables to 4 observers. Practice effects of two kinds occur, a short-term period of improvement when the crew tests unfamiliar conditions, and gradual long-term changes in skill. The first may be eliminated by repeating tests; the second may be evaluated by suitable control tests on reference circuits. The use of automatic equipment to facilitate the control of variable factors is very advantageous, particularly when considerable testing is done. Castner and Carter have described equipment of this kind. The sounds that the observer hears are recorded, and the recorded result is analyzed into various articulation values, automatically within a few seconds after the test is completed, by suitable selecting and recording equipment. In addition, various circuit conditions to be tested are set up at the proper time, the talking levels of the callers are measured, and numerous other steps in a test are accomplished automatically.

Another type of test that has been used somewhat is the "intelligibility test." In some such tests, short sentences of the interrogative or imperative form containing a simple idea are used. The sentences are considered to be understood if the observer either records the sentence or records an intelligent answer. In others, English words, selected at random from print, are used. The "discrete sentence intelligibility" is the percentage of the total number of called sentences which are correctly understood. Similarly, the "discrete word intelligibility" is the percentage of called words correctly observed. Figures 3 and 4 illustrate the types of relations between these intelligibilities and syllable articulation. The intelligibility tests are useful in testing very poor systems having an articulation of only a few per cent, which are used sometimes in fundamental studies.

Articulation and intelligibility tests afford a quantitative measure of the intelligibility of reproduced speech. The naturalness of speech and music, and the emotional and esthetic appeal of musical sounds, are less definite, and have been studied principally by means of the *AB* judgment test. In this test, the observer listens to two conditions, *A* and *B*, and judges which condition is distorted, or which condition is the poorer, etc., depending on the study. Sometimes one condition may be the original sound and the other condition the reproduced sound. Although such tests do not always give a quantitative measure of the effects of distortion on naturalness, esthetic appeal, etc., they do afford a means of measuring the just-noticeable amounts of distortion. This type of test was used in studies on the audible frequency ranges of music, speech, and noise which have been reported by Snow.

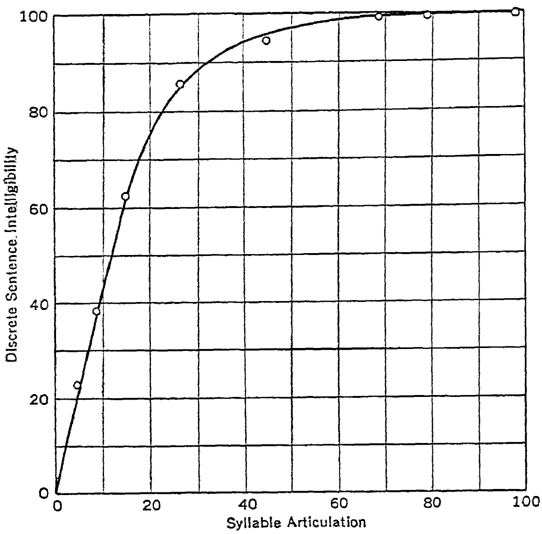


FIG. 3. Discrete Sentence Intelligibility vs. Syllable Articulation

FIELD TESTS. The results of such laboratory tests cannot be taken to indicate directly the performance of actual circuits under service conditions. In service, there is a wide range of variation in a number of conditions, such as the subject matter of conversations, the manner in which the transmitter is spoken into, the way in which the receiver is held to the ear, and the reactions of telephone users to the various circuit conditions. Because it is difficult to specify definitely such conditions and hence simulate them in the laboratory, the repetition test, which has been described by Martin, has come into use. In this test, observations are made on telephone circuits under service conditions. Essentially, the test is that of noting the repetitions in a telephone conversation and from a number of observations determining the repetition rate. Martin's paper describes repetition observations on conversations between several hundred stations in the American Telephone and Telegraph Company headquarters building and a similar number of stations in the Bell Telephone Laboratories building. To provide data for rating transmission performance, the station instruments and the transmission characteristics of the interconnecting trunks were varied and the effects on repetition rate determined. McKown and Emling have described a system of effective transmission data for rating telephone circuits, based on the results of such tests and articulation tests

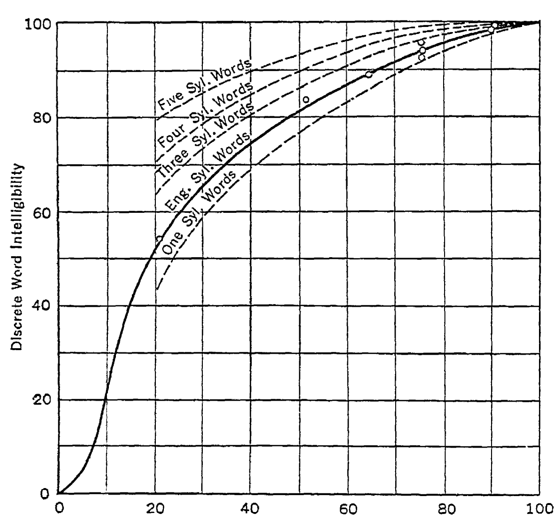


FIG. 4. Discrete Word Intelligibility vs. Syllable Articulation

OTHER TESTS. During the war a number of new tests were devised for testing speech transmission under conditions obtaining in military operations. See: *Transmission and Reception of Sounds under Combat Conditions*, Miller and Weiner, Columbia University Press, New York, 1948.

PREDICTION OF ARTICULATION TEST RESULTS. On many types of speech-transmission circuits it is possible to measure the circuit characteristics and noise levels and by means of these data compute the articulation score that would be obtained. Computational methods are also useful in the design of equipment for speech circuits and of considerable theoretical interest in the study of the interpretation of speech sounds. See: Beranek, L. L., Design of Speech Communication Systems, *Proc. IRE*, Vol. 35, 80 (1947); French and Steinberg, Factors Governing Intelligibility of Speech Sounds, *J. Acous. Soc. Am.*, Vol. 19, 90 (1947); and Fletcher and Galt, The Perception of Speech and Its Relation to Telephony, *J. Acous. Soc. Am.*, July 1950.

BIBLIOGRAPHY

Bureau of Publications, Bell Telephone Laboratories, High Speed Motion Pictures of the Human Vocal Cords. (Two reels of slow-motion views available for loan on request.)
Castner and Carter, Developments in the Application of Articulation Testing, *Bell Sys. Tech. J.*, July 1933, p. 347.
Crandall, I. B., Sounds of Speech, *Bell Sys. Tech. J.*, October 1925, p. 586.
Dudley, H., Remaking Speech, *J. Acous. Soc. Am.*, Vol. 11, 169 (October 1939).
Dudley, H., The Carrier Nature of Speech, *Bell Sys. Tech. J.*, Vol. 19, 495 (October 1940).
Dudley, Homer, and Otto O. Gruenz, Jr., Visible Speech Translators with External Phosphors, *J. Acous. Soc. Am.*, Vol. 18, 62 (July 1946).
Dunn, H. K., and D. W. Farnsworth, Exploration of Pressure Field Around the Human Head during Speech, *J. Acous. Soc. Am.*, Vol. 10, 184 (January 1939).
Dunn, H. K., and S. D. White, Statistical Measurements on Conversational Speech, *J. Acous. Soc. Am.*, Vol. 11, 278 (January 1940).
Firestone, F. A., Artificial Larynx for Speaking and Choral Singing, *J. Acous. Soc. Am.*, January 1940, p. 11.
Fletcher, H., Physical Characteristics of Speech and Music, *Bell Sys. Tech. J.*, July 1931, p. 349.
Fletcher and Steinberg, Articulation Testing Methods, *Bell Sys. Tech. J.*, October 1929, p. 806.
French, Carter, and Koenig, Words and Sounds of Telephone Conversation, *Bell Sys. Tech. J.*, April 1930, p. 290.
Koenig, W., H. K. Dunn, and L. Y. Lacy, Sound Spectrograph, *J. Acous. Soc. Am.*, Vol. 18, 19 (July 1946).
Kopp, G. A., and H. C. Green, Basic Phonetic Principles of Visible Speech, *J. Acous. Soc. Am.*, Vol. 18, 62 (July 1946).
Lewis, Don, Vocal Resonance, *J. Acous. Soc. Am.*, Vol. 8, 91 (October 1936).
Lewis and Tuthill, Resonant Frequencies and Damping Constants, *J. Acous. Soc. Am.*, Vol. 11, 451 (April 1940).
Martin, W. H., Rating the Transmission Performance of Telephone Circuits, *Bell Sys. Tech. J.*, January 1931, p. 116.
McKown and Emling, A System of Effective Transmission Data for Rating Telephone Circuits, *Bell Sys. Tech. J.*, July 1933, p. 331.
Paget, Sir Richard, Production of Artificial Vowel Sounds, *Proc. Roy. Soc.*, Vol. 102, 752 (1923).
Potter, R. K., Visible Patterns of Sound, *Science*, Vol. 102, 463 (Nov. 9, 1945).
Potter, R. K., G. A. Kopp, and H. C. Green, *Visible Speech.* Van Nostrand (1947).
Riesz, R. R., An Artificial Larynx, *J. Acous. Soc. Am.*, January 1930, p. 273.
Riesz, R. R., and L. Schott, Visible Speech Cathode Ray Translator, *J. Acous. Soc. Am.*, Vol. 18, 50 (July 1946).
Sacia and Beck, Power of Fundamental Speech Sounds, *Bell Sys. Tech. J.*, July 1926, p. 393.
Sivian, L. J., Speech Power and Its Measurement, *Bell Sys. Tech. J.*, October 1929, p. 646.
Sivian, Sunn, and White, Absolute Amplitudes and Spectra of Musical Instruments and Orchestras, *J. Acous. Soc. Am.*, January 1931, p. 330.
Steinberg, Application of Sound-measuring Instruments to the Study of Phonetic Problems, *J. Acous. Soc. Am.*, July 1934, p. 16.
Steinberg, J. C., and N. R. French, The Portrayal of Visible Speech, *J. Acous. Soc. Am.*, Vol. 18, 4 (July 1946).
Stewart, J. Q., An Electrical Analogue of the Vocal Organs, *Nature*, September 1922, p. 311.
Wolf, Stanley, and Sette, Quantitative Study of the Singing Voice, *J. Acous. Soc. Am.*, Vol. 6, 255 (April 1935).

EFFECTS OF DISTORTION ON SPEECH AND MUSIC

By John C. Steinberg and W. A. Munson

In studies of the effects of distortion, it is desirable to have a suitable reference system. Under certain conditions, air transmission from speaker to observer approaches such a reference. Owing to practical difficulties in changing and controlling an air transmission system, it is convenient to use electrical systems composed of transmitters, amplifiers, and receivers or loudspeakers. By suitable calibration and equalization, such systems can be made to approach air in transmission properties, and various amounts of degradation or improvement can be introduced. Circuits of this type have been studied by means of articulation and judgment tests and the effects of various kinds of distortion determined.

13. EFFECT OF FREQUENCY DISTORTION

AUDIBLE FREQUENCY RANGES OF MUSIC, SPEECH, AND NOISE. One step in the evaluation of the effects of distortion on the interpretation of sounds is the determination of the just-noticeable amounts of distortion. Snow has reported the results of tests on the lower and upper frequency limits of various types of sounds. The sounds as produced in one room were picked up by a microphone and transmitted to a second room where they were reproduced by means of loudspeakers. By means of filters, all frequencies above or below any desired cutoff could be suppressed. By comparing the unfiltered with the filtered transmission, trained observers determined the just-detectable upper and lower cutoffs. The results of the tests are shown in Fig. 1. The ends of the

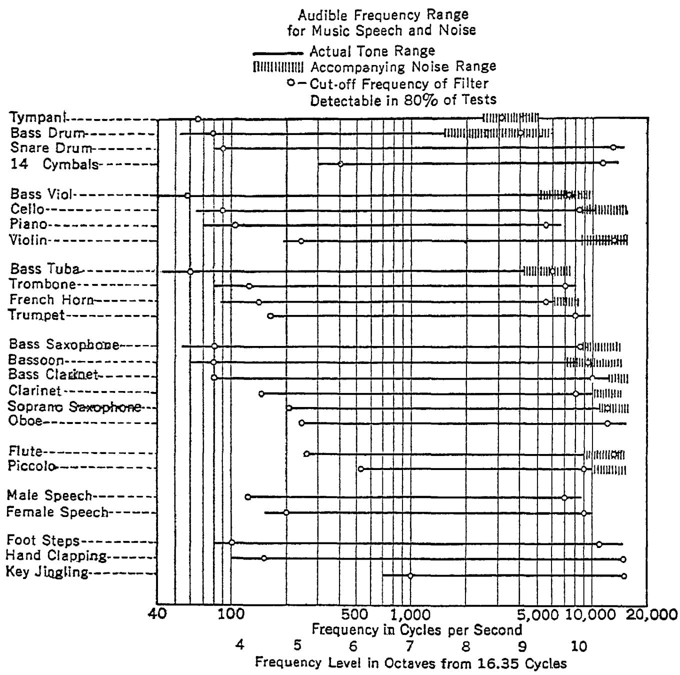


FIG. 1. Audible Frequency Ranges

lines representing the audible ranges are the cutoffs where the correct judgments amounted to 60 per cent. Circles mark the cutoffs where 80 per cent correct judgments were obtained. The musical sounds produced by many of the musical instruments were accompanied by noise such as key clicks, lip noises, "buzz" of reeds, and hissing of air. As far as possible, the observers tried to distinguish between tone and noise. The lines indicate their decision as to the actual tone range, and the short vertical lines indicate the noise range. In all cases, the upper frequency limit exceeds very greatly the frequencies of the highest notes that the instruments were capable of producing. Except for the piano, the lower limit was always set up by the frequency of the lowest note that could be produced.

EFFECT OF CUTOFF FREQUENCY ON ORCHESTRAL QUALITY. An additional type of test was made with music furnished by an 18-piece orchestra. In this test, the filtered reproduction was always presented on the *B* condition, and the unfiltered on the *A* condition. The observers were asked to rate the quality of the *B* condition as a percentage of the *A* condition. A rating of less than 100 per cent indicated a degradation; a rating of more than 100 per cent indicated an improvement. The various cutoffs occurred on the *B* condition in random order. The results of this test are given in Fig. 2. The judgment of the ten observers taking part in this test was that the quality increased rapidly as the cutoff was extended upward to 8000 cycles or downward to 80 cycles.

The results, in view of the nature of the test, were surprisingly consistent. In only about 3 per cent of some 400 observations was the quality of a filtered system given a rating greater than 100 per cent, so that, in general, the reproduction of the full audible range was preferred.

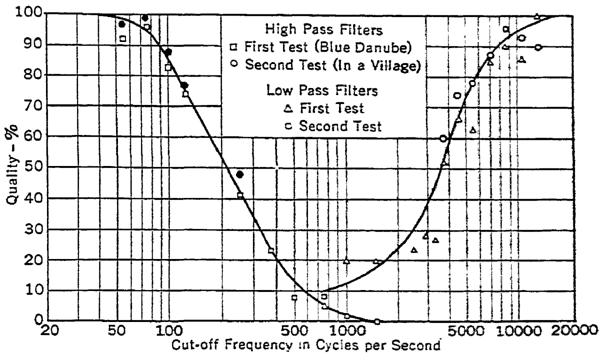


FIG. 2. Orchestral Quality vs. Cut-off Frequency

14. ARTICULATION TESTS

RELATION BETWEEN ARTICULATION AND LEVEL OF SPEECH ABOVE THRESHOLD. The articulation test affords a means of dealing with the effects of distortion in a more quantitative way than does the judgment test. An extensive series of articulation studies have been reported by Fletcher and Steinberg, for widely varying types of distortion. Figure 3 shows the effect, on syllable articulation, of changing the received speech level (see also article 31). In the range from 10 to 40 db above threshold, the articulation increases rapidly with increasing level. In the range from 50 to 90 db there is little change with increasing level. Experimentally it has been determined that, if a speaker radiates the average speech power of 10 microwatts, a listener of normal hear-

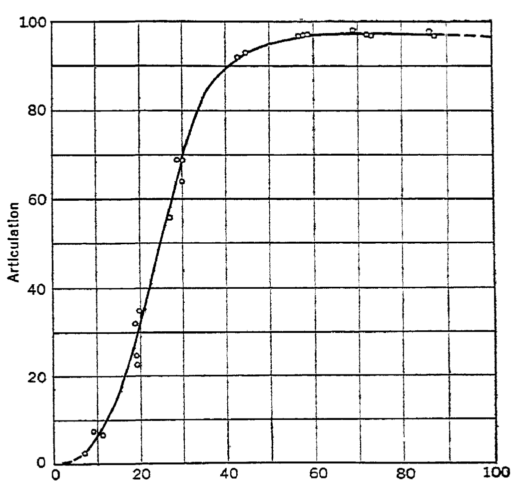


FIG. 3. Effects of Speech Level on Syllable Articulation

ing 1 meter away in the undisturbed sound field will hear the speech at a level of about 62 db above threshold. Figure 3 shows that this level is sufficient to insure very good articulation. It is a level that would obtain in conversing at a meter's distance in a quiet, well-damped room. A more detailed picture of the effects of speech level on articulation is obtained by grouping the sounds into five groups, where the sounds in each group have somewhat similar properties. The grouping used is: long vowels—*a*, *ā*, *ē*, *ō*, *o′*, *ū*; short vowels—*a′*, *e*, *i*, *o*, *u*, *w*, *y*; nasal consonants—*l*, *m*, *n*, *ng*, *r*; stop consonants—*b*, *ch*, *d*, *g*, *h*, *j*, *k*, *p*, *t*; fricative consonants—*f*, *s*, *sh*, *st*, *th*, *th′* (then), *v*, *z*, *zh* (see Table 1, p. 12-20). Figure 4 shows the group articulation vs. speech level. As the level is increased from threshold, the vowels are the first to be recognized and are followed in order by the nasal, stop, and fricative consonants. The apparent differences between the levels of the sound groups, as judged from articulation tests, correspond approximately with the differences in average phonetic powers shown in the insert. For example, the vowel articulation at a level of 25 db is about 90 per cent. The nasal consonant group must be increased 6 db from the 25-db level for an articulation of 90 per cent, and the stop and fricative consonants must be increased by 10 and 13 db, respectively.

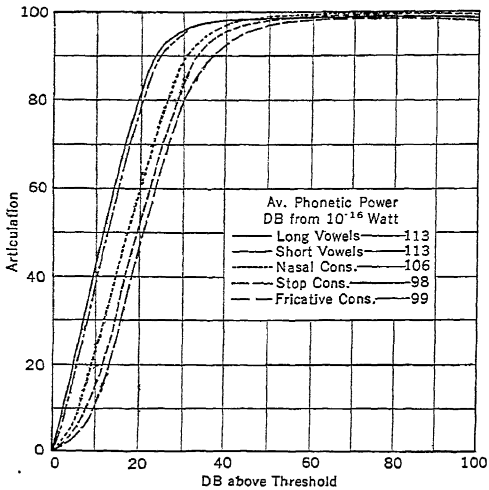


FIG. 4. Effects of Speech Level on Sound Articulation

EFFECT ON ARTICULATION OF REMOVING PORTIONS OF THE SPEECH FREQUENCY RANGE. To determine the effects of frequency distortion on articulation, electrical filters were introduced into a circuit that was otherwise reasonably free from distortion. One type of filter (low-pass filter) eliminated all components above a chosen cutoff frequency and uniformly transmitted the remainder of the range. Another type (high-pass filter) eliminated all components below a chosen cutoff frequency. Figure 5 shows the effects on articulation of changing the cutoff frequency of these filters. The level of the speech, before the filter was introduced, was held constant at 70 db above threshold. The curves marked L.P. are for the case where components below the cutoff frequency were transmitted; for the H.P. curves, components above the cutoff frequency were transmitted. The cutoff frequency at which the L.P. and H.P. curves intersect is

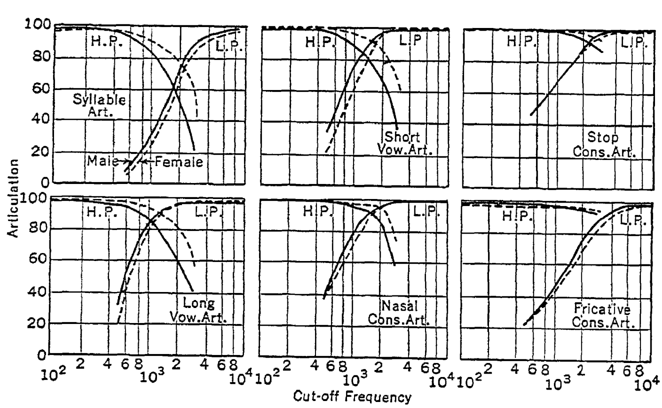


FIG. 5. Articulation vs. Cut-off Frequency

a frequency such that the frequency range below it has an articulation equal to the frequency range above it. This intersection frequency is always higher for female than for male speech, indicating that the higher-frequency ranges are more important for female speech. The frequency range in which the curves bend down is the range of importance in articulation, which as may be seen varies from one sound group to another.

EFFECTS OF AN EXTRANEOUS NOISE ON ARTICULATION. As pointed out in the article on masking, noise has the effect of shifting the threshold of audibility of other sounds heard in the presence of noise. If the masking of a noise is reasonably uniform over the important speech frequency range, articulation tests made in the presence of noise show that the main effect is to shift the curves of articulation vs. level above threshold. Figure 6 shows articulation vs. level for a quiet circuit and for the circuit with noise present at three different levels. The corresponding masking effects of the noise are shown by the insert (noise audiogram). The abscissa of the articulation vs. level curves gives the decibels above threshold of the speech on a quiet circuit. The amounts of the shift in the articulation curves correspond reasonably well with the average amounts of masking produced by the noise. When both the speech and noise levels are high, the

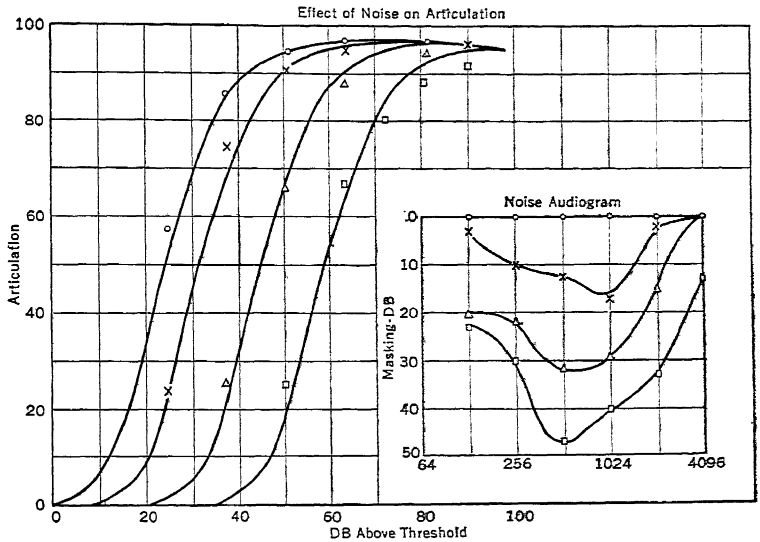


FIG. 6. Effects of Noise on Articulation

observed articulation values are less than would be indicated by a simple shift in the articulation-level curves. Modulation between the speech and the noise then takes place in the ear, which decreases the maximum value of articulation that can be obtained. When the masking is confined to a limited portion of the speech frequency range, such as that produced by a 2000-cycle tone, the effects on articulation cannot be represented by a simple shift in the articulation vs. level curves. In this case, the curves are both shifted and changed in shape. Moderate amounts of deafness affect articulation in much the same way as noise. When the hearing loss is uniform with frequency, the articulation to be expected is that obtained by a corresponding shift in the normal articulation vs. level curve.

EFFECT OF RESONANCE TYPE OF FREQUENCY DISTORTION ON ARTICULATION. A very common type of distortion occurring in transmission systems is that of resonance. Figure 7 shows the effects of a resonant peak at 1100 cycles, on articulation. The left-hand curves, numbered 2 and 3, show the loss in decibels at each frequency relative to the loss in the uniform system designated as 1. The right-hand curves show articulation vs. level for the three systems. The abscissa is the decibels above threshold of the normal speech, i.e., before inserting the resonance distortion. At the lower speech levels the decrease in articulation due to the distortion is much greater. Most of the articulation loss for a resonance type of distortion appears to arise because those sounds which have their important components in the frequency ranges removed from resonance are received at effectively lower than normal levels. For example, a large part of the important frequency range for fricative consonants is received at a level some 30 to 40 db below normal in curve 3. When the normal level is low, this further reduction causes appreciable articulation losses. Although the articulation at optimum received levels may not be seriously affected by resonance distortion, the tonal character or naturalness of the distorted speech is appreciably impaired.

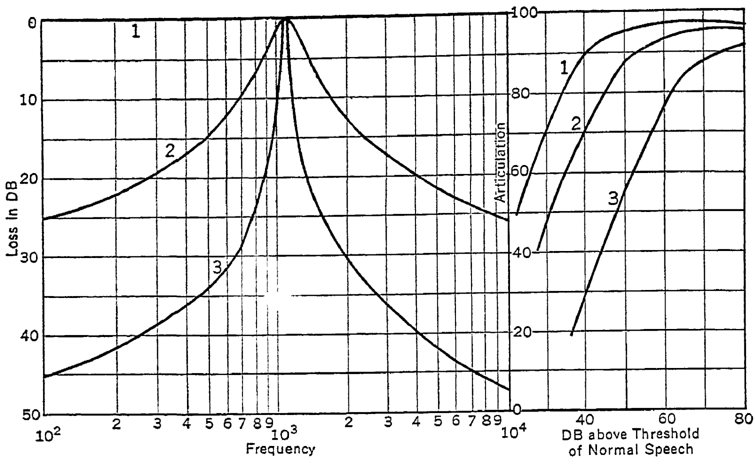


Fig. 7. Effects of Resonance Type of Frequency Distortion on Articulation

EFFECTS OF NON-LINEAR DISTORTION ON ARTICULATION. Non-linear distortion arises when the output of a system is not proportional to the input. There are two cases of interest: (*a*) when the output increases at a more rapid rate than the input; (*b*) when the output increases at a slower rate than the input. In both, modulation products or extraneous frequencies are introduced for the intense sounds when they cover an appreciable part of the curved portion of the input vs. output characteristic. The result is a decrease in the articulation of such sounds. The amount of the decrease depends upon the amount of the curvature. In addition, in (*a*) the intense sounds are amplified relative to the faint sounds, which has the effect of decreasing the articulation of the faint sounds. In (*b*) the faint sounds are amplified relative to the intense sounds. This has the effect of increasing the articulation and may, if the received speech level is below the optimum receiving level, offset the decrease due to modulation effects. For optimum receiving levels, however, the net result is always a decrease in articulation when the system is free from noise.

EFFECTS OF CLIPPING ON ARTICULATION. Articulation tests that have been made with voice-operated relays give an indication of the importance to articulation of

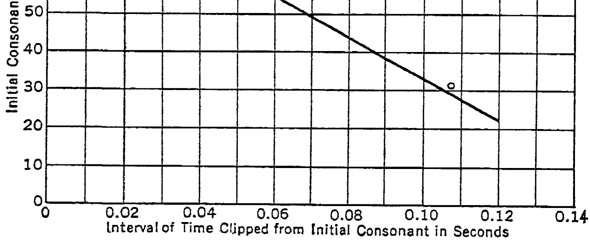


Fig. 8. Effect of Clipping an Element of Time from Speech Sounds

initial portions of the duration intervals of speech sounds. In the tests, syllables of the consonant-vowel-consonant type were spoken at intervals of about 3 sec. A circuit having a relay adjusted so as to break contact almost simultaneously with the beginning of a syllable was used. The contacts of a second relay formed a short circuit across the receiver. The operation of the first relay caused the second relay to break contact after an interval of time depending on the time constants of the relay circuit alone. The time taken for the second relay to operate represents the time clipped from the initial consonants of the syllables. Figure 8 shows the initial consonant articulation plotted against the operating time of the second relay. The data indicate that equal elements of time in the duration intervals of the sounds were of equal importance to the average articulation. For example, the average duration time of the initial consonants in these tests was about 0.16 sec. When a time interval of 0.08 sec was clipped from the sounds, the articulation was decreased by a factor of about 50 per cent.

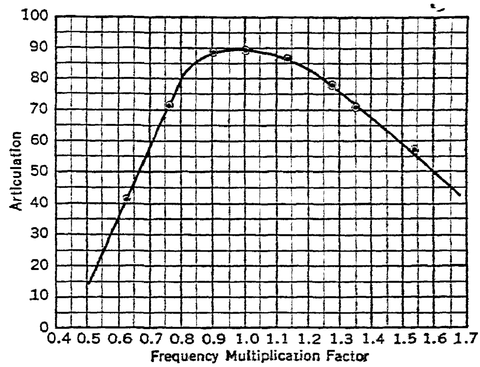


FIG. 9. Effect on Articulation of Multiplying Frequencies by a Constant Factor

EFFECTS OF FREQUENCY SHIFT ON ARTICULATION. One type of frequency shift, namely, the multiplication of frequencies by a constant factor, occurs when the speed of a sound film or phonograph turntable during reproduction is different from the speed used during recording. The effect of this type of distortion on articulation is shown in Fig. 9. Another type, the addition or subtraction of a constant number of cycles to all frequencies, occurs when the frequency of the carrier at the modulating end differs from that at the demodulating end of a carrier system. The effect of this type of shift is shown in Fig. 10. In general, the decrease in articulation is greater when the frequencies are shifted to lower values than when they are shifted toward higher values. In the first type of distortion the duration of the sounds is changed, whereas in the second type the duration is unchanged. In the first type the harmonic relationship is maintained, whereas in the second type this relationship is not preserved. In order to interpret speech sounds perfectly it seems to be necessary to preserve both these properties.

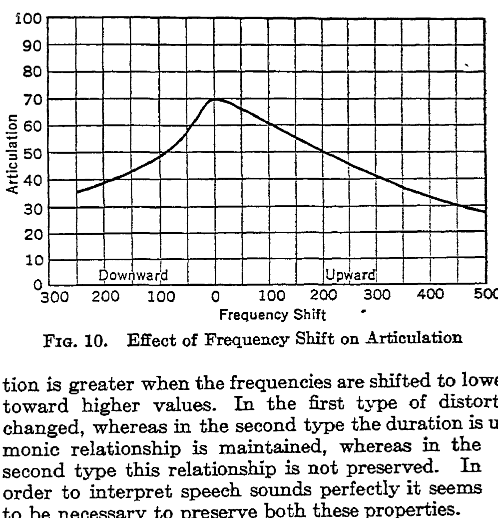


FIG. 10. Effect of Frequency Shift on Articulation

EFFECTS OF PHASE DISTORTION ON ARTICULATION. The aspect of phase distortion that has received most attention is the so-called "delay distortion" which arises when the phase characteristic of a transmission system is not proportional to frequency but is of the type shown by the curved line in Fig. 11. This type is of interest because it occurs in loaded lines and in low-pass filters. The delay distortion (see Section 5, article 9) is the difference between the slope of the phase characteristic at any frequency f and the minimum slope, and is usually expressed in the form

$$\left(\frac{\partial\beta}{\partial\omega}\right)_f - \left(\frac{\partial\beta}{\partial\omega}\right)_{\min}$$

where β is the phase shift in radians and $\omega = 2\pi f$. The approximate effect of this type of distortion is

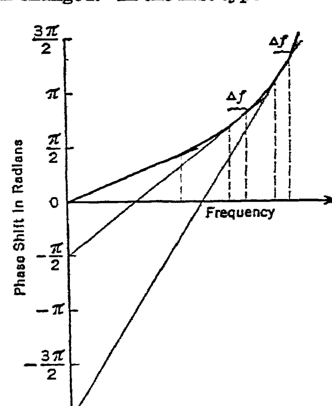


FIG. 11. A Curved Phase Characteristic

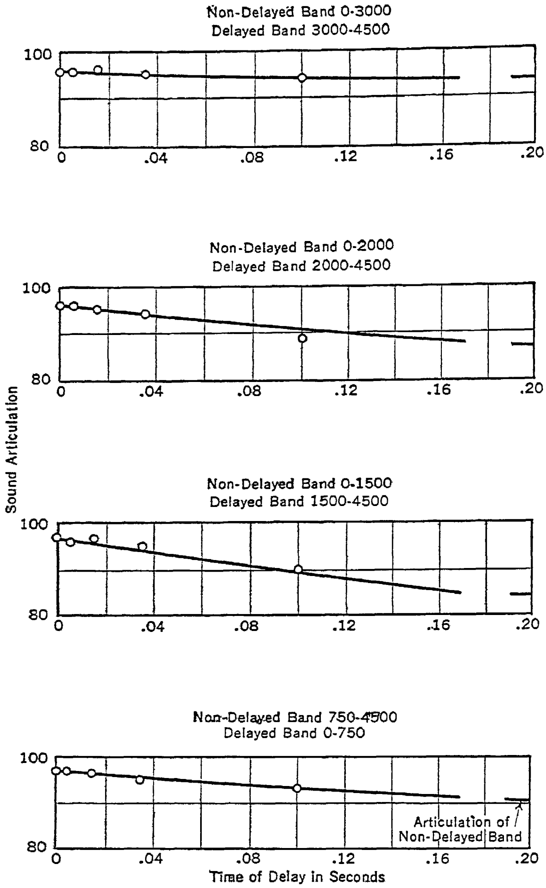


FIG. 12. Effects of Delay on Articulation

to delay the frequency components near f, with respect to those components in the range of minimum slope, by the amount of the delay distortion. The effects on articulation of delaying various portions of the speech frequency range are shown in Fig. 12. In these tests, a nominal undistorted speech frequency range of 0 to 4500 cycles was divided into two parts by means of filters and each part transmitted through a different channel. After transmission the two parts were recombined. The phase characteristic of each channel approximated a straight line over the greater part of the frequency range. The slope of the characteristic of one channel could be increased by various amounts over that of the other channel. One channel thus introduced a definite time delay, in the sense used here, with respect to the other channel, i.e., a delay given by the difference in the slopes of their phase characteristics. The articulation values decrease with increasing delay and approach the articulation of the more intelligible band which was also the least-delayed band.

To determine the effects on articulation of delay distortion which varied continuously with frequency, a system having the delay-distortion characteristics shown in Fig. 13 was studied. These characteristics were obtained from network sections of the all-pass type. By using different numbers of sections, different amounts of delay distortion could be obtained. The attenuation characteristics of the networks were equalized to 2500 cycles, and a 2400-cycle low-pass filter having negligible phase distortion was associated with the networks. Figure 14 shows the effect of the delay distortion on articulation.

To study the effects of delay distortion at higher frequencies, articulation tests were made on a system containing first one, and then twenty-five, 5000-cycle low-pass filters in series. In both cases, the attenuation was equalized to 5000 cycles. The delay-distortion characteristics and the results of

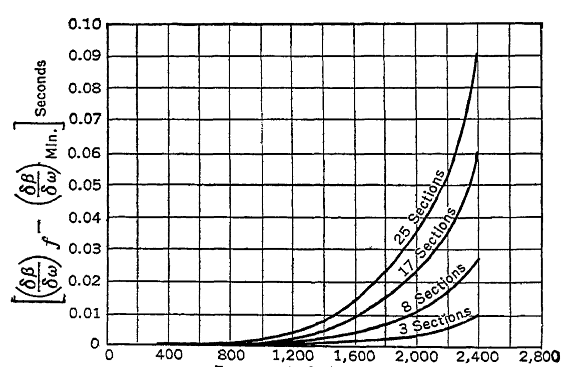


FIG. 13. Delay Distortion for an All-pass Network

the tests are shown in Fig. 15. Because the component frequencies near the cutoff are delayed by phase distortion of this type, they fail to contribute their normal amounts to the total articulation carried by the unattenuated frequency band. Thus the distortion has the effect of reducing the transmitted frequency range, so that the effective transmitting range depends upon both the phase and the attenuation characteristics.

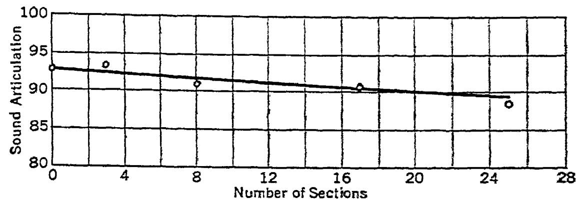


FIG. 14. Articulation vs. Delay Distortion

AUDIBLE EFFECTS OF PHASE DISTORTION. Quite aside from the effects on articulation, when speech from a system having phase distortion is compared with that from a system of negligible distortion, it is noticed that the distorted speech is accompanied by certain audible effects which appear to be extraneous to the speech and transient in character. These effects, which are due to components of different frequencies arriving at different times, are often termed "birdies" or "tweets," for delay distortion of the type shown above. Their noticeableness depends upon the amount of delay distortion and the frequency range in which it occurs. For speech, it was found that one section of the all-pass network, when associated with a 2400-cycle low-pass filter, had sufficiently small delay distortion so as to be just noticeable. This determination was made by alternately listening to speech from the system under two conditions: one, the filter alone; two, the

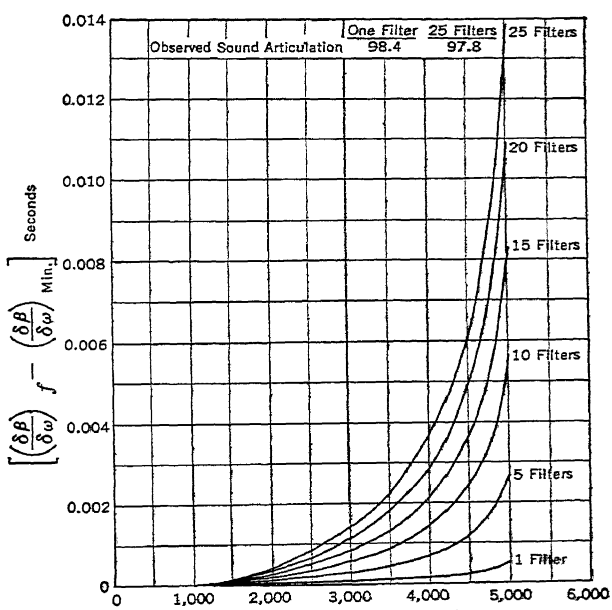


FIG. 15. Articulation and Delay Distortion for 5000-cycle Low-pass Filters

filter with the all-pass network. Judgments of which condition contained the network were correct about 50 per cent of the time and wrong about 50 per cent of the time for one section of the network. The total delay distortion at the cutoff frequency in this case, i.e., that due to the filter plus that due to one section of the network, was about 0.006 sec. When three sections were used the distortion was easily noticed.

Similar tests with the 5000-cycle low-pass filters indicated that some number between 5 and 10 filters in tandem would cause just-noticeable distortion and that the distortion was clearly noticeable for 20 filters in tandem. The amount of delay distortion at the cutoff frequency for 5 filters is about 0.003 sec, and for 10 filters about 0.006 sec.

The above figures depend somewhat upon the attenuation characteristic as the cutoff frequency is approached, small amounts of attenuation reducing the noticeability of the effects. The figures also vary somewhat with individuals, depending upon their experience and hearing characteristics.

Tests on piano reproduction when single notes were struck or when a passage of music was played indicated that the distortion caused by 25 of the 5000-cycle low-pass filters in tandem was not noticeable. As in speech, it would be expected that the noticeableness of the distortion would depend upon the frequency range in which it occurs. In general, the effects of delay distortion on music are very much less noticeable than on speech, which is probably due in part to the more sustained character of music.

EFFECTS OF ROOM REVERBERATION ON ARTICULATION. To determine the effects of room reverberation on articulation, the microphone of a system substantially

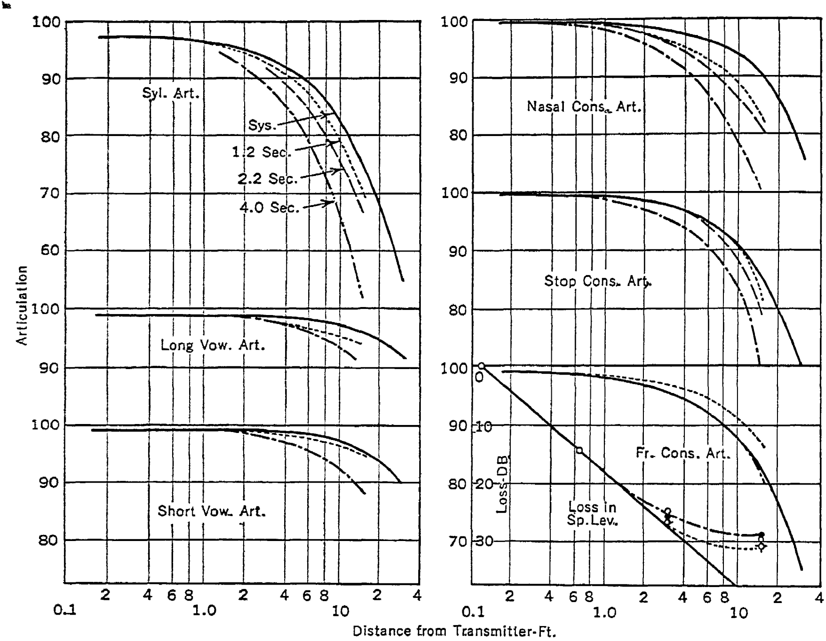


FIG. 16. Effect of Reverberation on Articulation

free of distortion was placed in a room having variable reverberation conditions. The room dimensions were 20 ft by 30 ft by 15 ft. The reverberation time was practically uniform over the important speech frequency range. The articulation tests were made under three conditions of reverberation, corresponding to the reverberation times 1.2, 2.2, and 4.0 sec, and for the following four distances between the diaphragm of the microphone and the lips of the speaker, 0.12 ft, 0.66 ft, 3.0 ft, and 15 ft. In all the tests, the speaker faced toward the microphone. Observers listened in another room, by means of head receivers, to the speech picked up by the microphone in the reverberant room. A volume indicator was used to measure the level of the speech from the microphone. The instrument measures, substantially, the phonetic power of the vowel sounds. The gain of the system between microphone and receiver was set so that the level of the speech received by the observers, for the close talking condition, was 74 db above threshold. The gain of the system was then kept constant for all tests. The loss in speech level vs. distance between microphone and speaker, and the articulation results, are shown in Fig. 16. The speech level decreased inversely as the square of the distance for distances up to 1 ft from the transmitter, but at the 15-ft distance the level was some 12 db greater than the level obtained by assuming the inverse square relation. The solid curves designated as "sys."

show articulation vs. distance between speaker and microphone, for the system alone. They are graphs of the data of Figs. 3 and 4, where level above threshold has been converted into distance between speaker and microphone. This was done by means of the solid curve showing loss in speech level vs. distance, and taking the level for zero loss as 74 db above threshold. This level is, of course, some 20 db less than the level the observer would have received if the speaker had spoken directly into his ear from a distance of 0.12 ft. A comparison of the curves indicates that the various sound groups were not uniformly attenuated by increasing the distance between speaker and microphone. For a given distance, the effect of increasing the reverberation time was always in the direction of decreasing the articulation, even though the received speech levels increased as the reverberation increased. For distances up to 15 ft, at least, the articulation was appreciably less for a reverberant room than it would be for one with perfectly absorbing walls.

15. AUDITORY PERSPECTIVE

Part of the emotional and esthetic appeal of sounds, when listened to directly with both ears, may be ascribed to the appreciation of their spatial character. When this property of sound waves is preserved in reproduction, the sounds are said to be reproduced in true auditory perspective. Ideally, there are two ways of accomplishing such reproduction. One is binaural reproduction, which aims to reproduce in the distant listener's ears, by means of head receivers, exact copies of the sound waves that would exist in his ears if he were listening directly. In this case, a complete system consisting of microphone, line, and receiver is used for each ear. The other method uses loudspeakers and aims to reproduce in the distant hall an exact copy of the pattern of sound vibration that exists in the original hall. In the limit, an infinite number of microphones and loudspeakers of infinitesimal dimensions would be needed. In a symposium on reproduction in auditory perspective the basic requirements for the production of orchestral music were discussed and a practical system designed to achieve auditory perspective was described (*Bell Sys. Tech. J.*, April 1934).

BIBLIOGRAPHY

Auditory Perspective, a collection of six papers, *Bell Sys. Tech. J.*, April 1934, p. 239.
Fletcher, An Acoustic Illusion Telephonically Achieved, *Bell Lab. Rec.*, June 1933, p. 286.
Galt, R. H., Noise Out-of-doors, *J. Acous Soc. Am.*, July 1930, p. 30.
Snow, W. B., Audible Frequency Ranges of Music, Speech, and Noise, *J. Acous. Soc. Am.*, July 1931, p. 155.
Steinberg, J. C., Effects of Distortion on the Recognition of Speech Sounds, *J. Acous. Soc. Am.*, October 1929, p. 121.
Steinberg, J. C., Effects of Phase Distortion on Telephone Quality, *Bell Sys. Tech. J.*, July 1930, p. 550.
Steinberg and Snow, Auditory Perspective and the Physical Factors Affecting It, *Bell Sys. Tech. J.*, April 1934, p. 245.
Tucker, R. S., Noise in Buildings, *J. Acous. Soc. Am.*, July 1930, p. 59.

ACOUSTIC PROPERTIES OF ROOMS

By Vern O. Knudsen

The acoustic quality of nearly all communication systems, as radio, television, telephony, motion pictures, phonograph recording, hearing aids, and public-address and other sound-amplifying or -reproducing systems, is largely influenced by the acoustic properties of the rooms in which the sound to be transmitted or recorded is generated, and in which this sound is reproduced.

The following definitions pertinent to the acoustic properties of rooms have been selected from American Standard Acoustical Terminology—Z24.1–1942 (see *J. Acous. Soc. Am.*, Vol. 14, 84–110 [1942] for these and other pertinent definitions).

Effective Sound Pressure (P). The effective sound pressure at a point is the root mean square value of the instantaneous sound pressure over a complete cycle, at that point. The unit is the dyne per square centimeter.

Pressure Level.* The pressure level, in decibels, of a sound is 20 times the logarithm to the base 10 of the ratio of the pressure P of this sound to the reference pressure P_0. Unless otherwise specified, the reference pressure is understood to be 0.0002 dyne per sq cm.

* In discussing sound measurements made with pressure or velocity microphones, especially in inclosures involving normal modes of vibration or in sound fields containing standing waves, caution must be observed in using the terms "intensity" and "intensity level." Under such conditions the terms "pressure level" or "velocity level" are preferable since the relationship between the intensity and the pressure or velocity is generally unknown.

Velocity Level.* The velocity level, in decibels, of a sound is 20 times the logarithm to the base 10 of the ratio of the particle velocity of the sound to the reference particle velocity. Unless otherwise specified the reference particle velocity is understood to be 5×10^{-6} cm per sec effective value.

Sound Energy Density (E). Sound energy density is the sound energy per unit volume. The unit is the erg per cubic centimeter.

Sound Intensity * (I). The sound intensity of a sound field in a specified direction at a point is the sound energy transmitted per unit of time in the specified direction through a unit area normal to this direction at the point. The unit is the erg per second per square centimeter, but sound intensity may also be expressed in watts per square centimeter.

Intensity Level * (I_L). The intensity level, in decibels, of a sound is 10 times the logarithm to the base 10 of the ratio of the intensity I of this sound to the reference intensity I_0. Unless otherwise specified the reference intensity I_0 shall be 10^{-16} watt per sq cm.

Echo. An echo is a wave which has been reflected or otherwise returned with sufficient magnitude and delay to be perceived in some manner as a wave distinct from that directly transmitted.

Multiple Echo. A multiple echo is a succession of separately distinguishable echoes from a single source.

Flutter Echo. A flutter echo is a rapid succession of reflected pulses resulting from a single initial pulse. If the flutter echo is periodic and if the frequency is in the audible range it is called a musical echo.

Noise. Noise is any undesired sound.

Acoustic Reflectivity. The acoustic reflectivity of a surface not a generator is the ratio of the rate of flow of sound energy reflected from the surface, on the side of incidence, to the incident rate of flow. Unless otherwise specified, all possible directions of incident flow are assumed to be equally probable. Also, unless otherwise stated, the values given apply to a portion of an infinite surface, thus eliminating edge effects.

Acoustic Absorptivity. The acoustic absorptivity of a surface is equal to 1 minus the reflectivity of that surface. (Sometimes called *Sound-absorption Coefficient.*)

Sabin. The sabin is a unit of equivalent absorption; it is equal to the equivalent absorption of 1 sq ft of a surface of unity absorptivity, i.e., of 1 sq ft of surface which absorbs all incident sound energy.

Acoustic Transmittivity. The acoustic transmittivity of an interface or septum is the ratio of the rate of flow of transmitted sound energy to the rate of incident flow. Unless otherwise specified, all directions of incident flow are assumed to be equally probable.

Reverberation. Reverberation is the persistence of sound, due to repeated reflections.

Rate of Decay (of Sound Energy Density). The rate of decay of sound energy density is the time rate at which the sound energy density is decreasing at a given point and at a given time. The practical unit is the decibel per second.

Reverberation Time (T). The reverberation time for a given frequency is the time required for the average sound energy density, initially in a steady state, to decrease, after the source is stopped, to one-millionth of its initial value. The unit is the second.

Mean Free Path. The mean free path for sound waves in an inclosure is the average distance sound travels in the inclosure between successive reflections.

16. REQUIREMENTS FOR GOOD ACOUSTICS

In order that a room may have good acoustics it is necessary (1) that the sound be sufficiently loud in the room; (2) that the room be free from noise whether of internal or external origin; (3) that the room be free from echoes, "flutters," or other interfering reflections; (4) that the reflecting boundaries of the room be so disposed as to provide a nearly uniform distribution of sound energy throughout the room; (5) that the room be free from undesirable resonance; and (6) that the reverberation in the room be sufficiently reduced to avoid excessive overlapping or commingling of successive sounds of articulated speech or music, but that the room be sufficiently "live" at all frequencies to give a pleasing effect to either speech or music as judged by the average listener.

In order to attain these necessary conditions for good acoustics the architect and acoustical engineer must assume responsibility relating to the following: (1) the selection of the

* In discussing sound measurements made with pressure or velocity microphones, especially in inclosures involving normal modes of vibration or in sound fields containing standing waves, caution must be observed in using the terms "intensity" and "intensity level." Under such conditions the terms "pressure level" or "velocity level" are preferable since the relationship between the intensity and the pressure or velocity is generally unknown.

site; (2) the making of a noise survey in the proximity of the proposed site; (3) the selection of a general type of wall and ceiling construction which will insulate the building adequately against external noise and vibration; (4) the selection and arrangement of rooms which require acoustical designing; (5) the design of the rough sketches for all speech rooms, music rooms, recording or broadcasting rooms, based upon the requirements for the proper distribution of direct and of reflected sound; (6) the application of appropriate formulas and principles to the detailed design of shape, sound insulation, and sound absorption for all rooms which require acoustical designing; (7) the selection of materials which will satisfy the acoustical, structural, decorative, and economic requirements; (8) the supervision of all aspects of construction which will affect the outcome in acoustics, and especially the making of tests on such materials as acoustical plaster; and (9) the testing of the completed building with regard to the distribution of sound; freedom from echoes, sound foci, or interfering reflections; the optimal conditions of reverberation; and the adequacy of sound insulation. In general, the acoustic problem consists of the adequate reduction of noise and vibration, and the designing of interiors in which the voice or instrumentation is heard or recorded most satisfactorily. All the factors mentioned in this section relate to the problem of room acoustics. One of the most basic of these factors is the problem of the growth and decay of sound in a room or, more strictly, the transient and steady-state behavior of sound in that room. Before considering this problem, it will be helpful to describe briefly two possible approaches to this and related problems: one, geometric acoustics, and the other, wave acoustics.

17. GEOMETRIC AND WAVE ACOUSTICS

Geometric or "ray" acoustics, as applied to the acoustics of rooms, assumes that sound travels in rays, that its frequency remains unchanged during the transient state as well as the steady state, that the rays are reflected, with partial absorption and transmission, at each encounter with the boundaries of the room, and that after a number of successive reflections the sound in the room becomes diffuse (all directions of propagation being equally probable) and of uniform energy density throughout the room. Obviously, this oversimplifies the actual behavior of sound waves in the room, especially when the wavelengths of the sound, as is often the case, are not small compared with the dimensions of the room; it neglects entirely such important properties as the normal modes of vibration of the room, interference, and diffraction. In spite of these shortcomings, it leads to many principles and formulas by means of which the acoustical engineer or architect can design auditoriums with satisfactory acoustics.

Wave acoustics deals with sound as waves; it offers the only means of dealing rigorously with wave phenomena in bounded spaces, with interference, diffraction, normal modes of vibration, and with the influence of localized areas of absorptive material and of irregularities of the contours of the boundaries of the room on both the transient and steady states of the sound in the room. The difficulties in applying wave acoustics to rooms are so great that very little progress has been made until recently, and even now much remains to be done before it can be used effectively by the average architect or acoustical engineer.

Historically, the study of the acoustic properties of rooms began and has continued, very largely, on the assumption that geometric acoustics is adequate to cope, at least approximately, with the transient and steady-state behavior of sound waves in most rooms. Obviously, the methods of geometric acoustics furnished satisfactory approximations in rooms in which the ratio of the room dimensions to the wavelength of the sound is large and in which the sound distribution is thoroughly diffuse (the "ergodic" state), but in most rooms, and especially for sounds of low frequency, the more rigorous methods of wave acoustics, recently developed by Morse and Bolt, should be used so far as this is possible. (See Morse and Bolt, Sound Waves in Rooms, *Reviews of Modern Physics*, April 1944). These methods of wave acoustics already have explained many apparent anomalies in room acoustics, especially those relating to the absorptive properties of acoustic materials as used in different rooms. No one who is not familiar with these methods should undertake the design of a room in which good acoustics is a prime requirement. Unfortunately, the results of these studies are not yet sufficiently simplified to be used by the average engineer as the basis for making routine analyses of the acoustic properties of rooms. Fortunately, on the other hand, the approximate and much simpler methods of geometric acoustics, when used with an understanding of the consequences and modifications which wave acoustics entails, will be found to be satisfactory for making the routine calculations which govern the acoustic properties of most rooms, such as offices, restaurants, classrooms, and residential rooms. For the acoustic design of broadcasting and motion-picture studios, auditoriums, theaters, churches, music rooms, courtrooms, lecture rooms, and all other rooms in which high-quality speech and music are required, full use should be made of the methods of wave acoustics, as far as they are applicable.

For the present, the methods of geometric acoustics continue to be used by most engineers for designing or correcting the acoustic properties of rooms, and their methods probably will continue to be used for possibly another five to ten years, but they should be used only in the light of existing and advancing knowledge of wave acoustics. In the following treatment, based largely on geometric acoustics, the practical modifications which wave acoustics imposes or suggests will be considered, and those who use this handbook as a guide for planning for good acoustics should give similar consideration to the relevance of wave acoustics in modifying the calculations and conclusions based on geometric acoustics.

18. GROWTH AND DECAY OF SOUND IN ROOMS—GENERAL CONSIDERATIONS

The approximate theories in use today are based upon the assumptions that sound, originating at some point in a room, propagates rays of vibratory energy with a speed of about 1125 ft per sec, uniformly in all directions; that these rays are partially reflected by the boundaries of the room; and that even after the source of sound is stopped these rays persist with their original frequency but become feebler after each reflection until ultimately they become inaudible. In these approximate theories it is assumed that the sound energy persists in rays or bundles; that, during the decay, the sound energy in the room remains constant in these rays or bundles for a short interval of time, equal to the time required for a ray to travel the average distance between successive reflections, the *mean free path*, and that then the total sound energy in the room suddenly drops a certain amount determined by the "average" (usually the weighted arithmetical mean) acoustic absorptivity of the boundaries of the room; and that this process of absorption by discrete steps continues until all the sound energy is converted into heat.

Although most of the absorption takes place at the boundaries at low frequencies, the absorption in the air at frequencies above 5000 cycles may be greater than the absorption at the boundaries. If the source continues to generate sound at a constant rate, a condition of equilibrium will be reached in which the rate of supply of sound energy to the room is just equal to the rate of absorption by the air and the boundaries. If the source is then stopped the sound in the room will die away at a rate equal to the rate of absorption, which is determined principally by the size, the shape, and the boundaries of the room. Although this decay is strictly made up of the free, damped, normal modes of vibration of a three-dimensional continuum, the decay is approximately represented by the simplification described above, provided that the absorptive material is distributed throughout the boundaries of the room and especially that it is not concentrated on one or two walls of the room. (Walls, as here used, refers also to the floor and ceiling.)

According to this simplification, the time required for the intensity of the sound to be reduced a specified amount will depend upon (1) the number of reflections which occur per unit time, and (2) the amount of sound energy which is absorbed at each reflection. If the room is a large one there will be only a few reflections per second; and in addition, if but a little sound energy is absorbed at each reflection, it will require a relatively long time for the intensity of ordinary sound to be reduced to the threshold of audibility. Such a room will be excessively reverberant. On the other hand, if the room is small and the boundaries highly absorptive, the room will be free from reverberation. Since the average intensity of speech or music in a room is of the order of one million times the intensity which is just barely audible, and since the hard, rigid boundaries may reflect as much as 98 per cent of the incident sound energy, it is apparent that an appreciable time, amounting to several seconds in many instances, is necessary for the sounds of speech or music to be reduced to inaudibility.

Thus, consider a room having a mean free path of 51 ft. Since the velocity of sound at room temperature (20 deg cent) is approximately 1122 ft per sec, there will be in this room just 22 reflections each second. Hence, if the initial sound in this room has an intensity of one million threshold units, and if 98 per cent of the initial sound energy is reflected at each encounter with the boundaries, to reduce the sound energy to one-millionth of its initial amount would require n successive reflections, where n is given by $0.98^n = 0.000001$. Solving, n is 684; that is, it requires 684 successive reflections in this room for the sound energy to die away to inaudibility. Since 22 reflections occur each second in this room, the time required for the sound to die away to inaudibility is $684 \div 22$, or 31.1 sec; that is, the time of reverberation in this room is 31.1 sec. By a similar consideration it can be shown that if this same room were completely lined with a material that reflects 50 per cent of the incident sound energy the total number of reflections would be reduced to 19.9, and the time of reverberation would be of the order of 0.9 sec. These simple considerations neglect the absorption of sound in the air, which at high frequencies will greatly modify these calculations. This type of absorption will be considered later in this section.

The formulas to which such approximate theories lead are sufficiently valid for most practical purposes in rooms which are bounded by materials having the same coefficient of absorption. However, in rooms bounded by materials having widely different coefficients of absorption, the formulas approach validity only as the decadent sound in the room is made to approach a completely diffuse state. It should be clearly recognized therefore that the formulas which will now be presented must be used with *caution* and *understanding*, especially with reference to the *average* coefficient of absorption in rooms which are bounded partly by highly reflective and partly by highly absorptive materials.

19. REVERBERATION EQUATIONS

The early experiments of W. C. Sabine show that the time of reverberation in a room is proportional directly to the volume of the room and inversely to the total amount of absorption supplied by the boundaries of the room. Sabine was able to determine experimentally the constant of proportionality k between reverberation time T and the volume V divided by total absorption a. Thus,

$$T = \frac{kV}{a} \tag{1}$$

The value of k as determined experimentally by Sabine for a large number of rooms of different shapes and sizes, at normal room temperatures, is approximately 0.05 when V is in cubic feet and a is in sabins (British units) and 0.164 when V is in cubic meters and a is in square meters (metric units). A few years later, Franklin and also Jaeger obtained this same equation from theoretical considerations.

This equation was used for nearly 30 years for calculating the reverberation time of either contemplated or finished rooms. Even the fallacious conclusion to which the equation leads for a room with totally absorptive surfaces, namely, that $T = kV/S$ instead of zero (where S is the total surface area of the room), was overlooked or was not sufficiently disturbing to destroy confidence in its validity, until recently. The equation is satisfactory in practice for frequencies between about 200 and 1000 cycles in the large majority of rooms in which the absorptive material is distributed and in which the rate of decay of sound is slow; that is, the equation applies to "live rooms," provided the frequencies are high enough to be well above the fundamental resonant frequency of the room but not high enough to involve a consideration of the attenuation in the medium.

MODIFICATION OF THE REVERBERATION FORMULA. A more satisfactory reverberation formula can be obtained by recognizing that the decay of sound takes place *discontinuously*, at time intervals equal to the time required for sound to travel the mean free path. The mean free path is given for rooms of conventional shape as $4V/S$. Strictly it is dependent upon the shape of the room and the location of the source, so that the equation is not accurate for rooms of peculiar shape. Thus, for the usual location of the source of sound and for the first 25 successive reflections, the mean free path is $4.3V/S$ for a large church or cathedral of cruciform shape and is $3.7V/S$ for a room with large horizontal dimensions and a low ceiling, the usual shape for large office and work rooms.

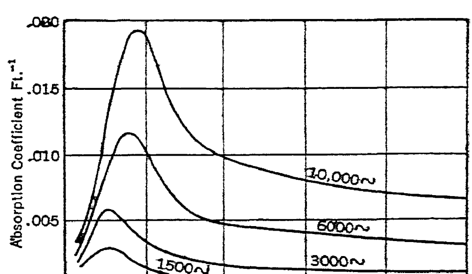


FIG. 1. Coefficients of Absorption of Sound in Air Containing Different Amounts of Water Vapor, for Frequencies of 1500, 3000, 6000, and 10,000 Cycles

Allowance must also be made for the absorption of sound in the air of the room, which is of considerable importance at the higher audible frequencies and especially in large rooms. The curves of Fig. 1 give the absorption coefficient m per foot for plane waves in air at 20 deg cent. It will be seen from the curves in Fig. 1 that m has a maximum for a certain concentration of water vapor, different for each frequency.

Based upon the inclusion of these factors and a mean free path of $4V/S$ the time of reverberation is

$$T = \frac{55.3V}{c[4mV - S \ln (1 - \bar{\alpha})]} \tag{2}$$

where c is the velocity of sound and $\bar{\alpha}$ is the arithmetic mean of the absorption coefficients of all the boundaries of the room.

At room temperature, 21 deg cent, $c = 1125$ ft per sec, so that for most working conditions

$$T = \frac{0.049V}{4mV - S \ln (1 - \bar{\alpha})} \tag{3}$$

in British units, or

$$T = \frac{0.161V}{4mV - S \ln (1 - \bar{\alpha})} \tag{4}$$

in metric units (see above). For frequencies below about 1000 cycles, m is so small that the first term in the denominator of eqs. (2), (3), or (4) can be neglected, that is, the absorption in the air is inappreciable; whereas, at high frequencies (above 5000 cycles), this term may become larger than the second (or surface absorption) term. At sufficiently high frequencies (above the audible range), the second term will become negligible, in which case the rate of decay and consequently the time of reverberation will be independent of the size of the room.

The foregoing reverberation formula, eq. (2), is sufficiently valid for practical purposes provided the sound in the room is thoroughly diffuse throughout the decay. This condition is realized for frequencies above about 250 cycles in all but very small rooms, provided all the boundaries of the room have approximately the same absorptivity, or provided suitable rotating paddles or "warble tones" are used to "mix" the sound in the room. Suitable precautions, such as those just mentioned, can be taken in making measurements in an acoustical laboratory, such as a reverberation chamber. In many rooms encountered in practice, the absorptive material may be concentrated on a single surface, as when a carpet, upholstered seats, and audience are all located on the floor, and the other surfaces in the room are highly reflective. In such rooms, especially if the opposite walls are parallel and not too far apart, the decay of sound will not conform to the approximately exponential decay predicted by eq. (2) but will consist, first, of a rapid rate of decay while the sound is relatively diffuse, and second, of a much slower rate of decay, made up largely of a horizontal flow of sound energy between the parallel and highly reflective walls, i.e., of modes of vibration which are directed at grazing incidence to the floor. The time of reverberation in such a room will be longer than that calculated by means of eq. (2), using an arithmetical mean for $\bar{\alpha}$.

This is borne out by some oscillograms of the rate of decay obtained in a small room, 8 ft by 8 ft by 9.5 ft (high), with the floor covered with a material having a rated absorption coefficient of 0.60 at 512 cycles and with the walls and ceiling finished with painted concrete. The first part of the decay (15 to 17 db) was relatively rapid, 95 to 100 db per sec. This was followed by a much slower decay, 38 to 40 db per sec. If the first part of the decay is used for calculating the time of reverberation T and the absorptivity of the floor material α, we obtain $T = 0.61$ sec and $\alpha = 0.55$. If the latter part is used, $T = 1.54$ sec and $\alpha = 0.22$. It will be noted that the first part of the decay, while the sound is relatively diffuse, yields a value for α which agrees fairly well with the rated value of 0.60 and therefore conforms reasonably well with the requirements of eq. (2), whereas the latter part of the decay, which is made up largely of the horizontal modes of vibration, is much slower than would be predicted by eq. (2). This is an extreme example of the inadequacy of eq. (2), which is based on geometric acoustics, to account for the true nature of the decay of sound in a room in which the absorptive material is concentrated on one wall (or floor or ceiling). Wave acoustics, on the other hand, accounts for the observed results very satisfactorily.

Fortunately, for the best acoustical quality in a room the absorptive material should be distributed on all surfaces of a room so that the rate of decay will be at least approximately the same in all directions, and under these circumstances eq. (2) will yield results which, as a rule, do not differ more than 10 per cent from the observed values. Furthermore, in very large rooms, as in theaters, school auditoriums, and churches, there is very little tendency for the reverberation to persist in two dimensions, even though most of the absorption is concentrated on the floor or on the floor and in the ceiling, (1) because the dimensions of the room are large compared with the wavelengths of the sound, and (2) because the architectural treatment of large rooms usually involves structural forms and ornamentations which tend to diffuse the sound during free decay.

In such rooms, provided there are no curved surfaces giving rise to concentrations of sound, the first 30 db (or more) of decay conforms very closely to eq. (2); and it is this portion of the decay that is pertinent to the acoustic quality of speech and music in rooms. Stated otherwise, the rate of decay after the first 30 db of decay is of little consequence, since in articulated speech or music such residual sounds will be so weak as to be completely masked by the primary (and much louder) sounds that follow. It is apparent, therefore, that eq. (2) is satisfactorily valid for the practical calculations of reverberations in most rooms. In many small rooms, such as are frequently used for radio broadcasting or for the recording of sound, it is important that the sound-absorptive materials be distributed quite uniformly throughout the room if eq. (2) is to be applicable. The true rate of decay of sound in small rooms is greatly dependent upon room resonance.

20. ROOM RESONANCE

A room is a resonant chamber, with resonant properties similar to those of a violin string, an organ pipe, or a diaphragm of a telephone receiver, except that in general the resonant properties of the room are much more complicated than those of one- or two-dimensional systems. Thus, if the dimensions of a rectangular room are l_1, l_2, and l_3, the "resonant" or normal frequencies ν for the room are given approximately by

$$\nu = \frac{c}{2}\left[\frac{l_1^2}{n_1^2} + \frac{l_2^2}{n_2^2} + \frac{l_3^2}{n_3^2}\right]^{1/2} \tag{5}$$

where c is the velocity of sound in the room, and n_1, n_2, and n_3 are order numbers having values of 0, 1, 2, 3, ... In a room where $l_1 = 8$ ft, $l_2 = 8$ ft, and $l_3 = 9.5$ ft, the gravest mode of vibration is given by $n_1 = n_2 = 0$ and $n_3 = 1$; that is, the room resonates as does an organ pipe, 9.5 ft long and stopped at both ends, when vibrating at its gravest mode. The wavelength of this gravest normal mode is 19 ft, and the frequency (assuming $c =$ 1125 ft per sec) is 59.2 cycles. Other resonant or normal frequencies, in ascending order, for this room, are 70.3, 92.8, 99.8, 116.0, 119.2 cycles, and continuing in a triply infinite series of frequencies corresponding to increasing values of the n's. When none of the n's is zero the resonant standing waves are *oblique*, and when one or two of the n's are zero the waves are *tangential*; that is, the waves move parallel and grazing to one or two, respectively, pairs of walls in the rectangular room.

The above-enumerated normal frequencies for the room under discussion have been observed, both by the "reinforcement" the room gives to steady tones of these frequencies, and by the transient decay of tones which have approximately these frequencies. Thus, Fig. 2 is a series of oscillograms showing the free decay of sound for this room when it is excited with tones having different "driving" frequencies between 90 and 101 cycles, as indicated above each oscillogram. When the room is excited with a frequency of 92.9 cycles, which corresponds closely to the normal mode having a frequency of 92.8 cycles, the free decay is made up almost exclusively of this one mode, and its decay rate is smooth and almost exactly exponential. Similarly, when the room is excited by any other frequency which differs only slightly from 92.8 cycles, the free decay is made up of this one mode, and the frequency during decay *is not the driving frequency* but is *always the frequency of this normal mode.* When the room is excited with a frequency of 96.7 cycles, which is about midway between the frequencies of the two modes of 92.8 and 99.8 cycles, the free decay is made up of these two normal modes, which are about equally excited, giving rise to the pronounced beats of about 7 per second, as expected, since the two natural frequencies differ by 7 cycles. When this same room was excited at a frequency of 118 cycles, the oscillogram of the free decay revealed the coexistence of beats at about 3.3 and 19.3 cycles, which no doubt arose from the simultaneous excitation of the three neighboring normal frequencies of 99.8, 116.0, and 119.2 cycles. No matter what frequency is used for exciting the room, the "reverberation" always consists of the free decay of one or more of the room's normal modes of vibration.

The foregoing results demonstrate that reverberation is not the persistence of *rays* of sound which continue, after the source is stopped, as rays of sound successively reflected back and forth in the room, but rather is the free decay of one or more (and usually many) modes of vibration. However, Strutt has shown that the formal law governing the free, damped vibration of the sound in a room approaches asymtotically the simple reverberation law of eq. (2) as the wavelength of the exciting sound becomes short in comparison with the wavelength of the gravest mode of vibration for the room. Thus, in a room (assuming the absorptive material to be fairly well distributed) having its longest dimension 10 ft, which is about the smallest room in which acoustics is a factor of importance, the wavelength of the gravest mode is 20 ft. In such a room, sound having a wavelength of one-tenth of that of the fundamental, namely, 2.0 ft, would be sufficiently short to conform reasonably well to eq. (2).

In other words, when the room is filled with sound having a wavelength shorter than 2.0 ft, that is, a frequency greater than about 560 cycles, the modes of vibration which are excited are so numerous and their frequencies so close together that the sound in the room is essentially diffuse and, therefore, the requirements are fulfilled for the approximate theories of reverberation described in the earlier paragraphs of this section, especially if

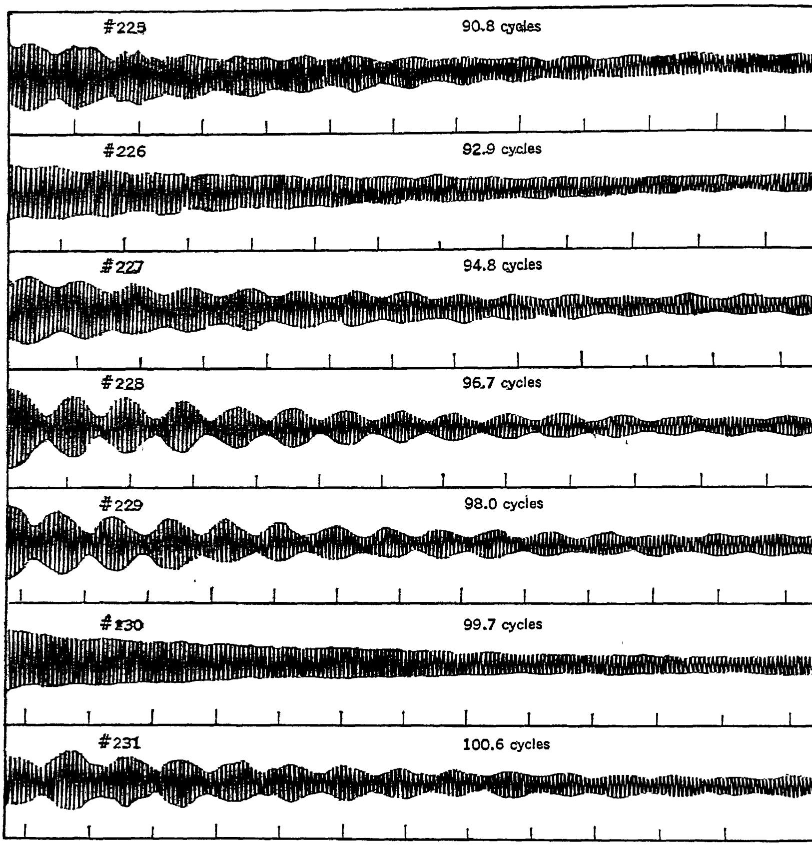


FIG. 2. Oscillograms of the Decay of Sound in a Small Rectangular Room, Showing that the Decay of Sound Consists of the Damped Free or Normal Vibrations of the Room

the absorptive material is distributed uniformly over the boundaries of the room or if some means are provided for mixing or diffusing the sound during the decay. In large rooms, such as concert halls, church or school auditoriums, and theaters, the lowest modes of vibration are usually in the subaudible range of frequencies, so that the elementary theory of reverberation applies with adequate rigor in such rooms for all frequencies above about 100 cycles, and the effects of room resonance usually can be neglected. (For further details consult M.J.O. Strutt, *Zeitschrift angew. Math. Mech.*; Knudsen, *J. Acous. Soc. Am.*, Vol. 4, 20–37 [1932]; and Morse and Bolt, *loc. cit.*)

REVERBERATION IN COUPLED SPACES. When two or more inclosures are coupled by means of openings such as open doors, passageways, or even thin partitions which are capable of transmitting an appreciable amount of sound, the reverberation in each inclosure is affected by the reverberatory properties of the other inclosure. Thus, most auditoriums of the theater type are divided into at least three coupled spaces—the stage, the main portion of the auditorium, and the space under the balcony. Even a long sound-recording studio, one end of which is reverberant and the other end non-reverberant, may be regarded as two coupled spaces. Office space is often divided into a number of coupled spaces; and a great complexity of coupled spaces often will be found in cathedrals, consisting of nave, transepts, choir, sanctuary, aisles, chapels, balconies, and organ chamber. If the mean free path can be determined for such coupled spaces, and if all surfaces have approximately the same absorption coefficients, the regular reverberation formula, using the appropriate value of k, will give the time of reverberation for the entire room. (See Knudsen, *Architectural Acoustics*, Chapter V.) But it is not always feasible to determine the mean free path for a complicated combination of coupled spaces, and it is very improbable that all surfaces will have even approximately the same coefficients of absorption. In many instances, therefore, it becomes necessary to consider the reverberation in each of the several coupled spaces, and to adjust the reverberation time in each space to the optimal condition.

Many rooms have poor acoustics because of failure to recognize the effect of coupled spaces. Thus, in many school auditoriums containing a balcony, it is common practice to install nearly all the absorptive material in the ceiling. In some auditoriums the walls under the balcony, the soffit of the balcony, and the floor under the balcony may be of hard, reflective materials, as hard plaster and concrete. If, in addition, the seats under the balcony are of the unupholstered type, the space under the balcony will be very reverberant although the space in the main part of the auditorium may be quite free from reverberation. During the growth or decay of sound in such an auditorium there is a transfer of energy between the two coupled spaces, with different rates of growth or decay in the two spaces. During the steady state the rate of transfer of sound from the *dead* space to the *live* one is equal to the rate of transfer in the opposite direction. During the very early stages of the decay these rates of transfer are nearly equal, but since the sound decays much more rapidly in the main part of the auditorium than it does under the balcony, there soon will be established an excess rate of flow from the live to the dead space, and the result is that the reverberation is prolonged in the main part of the auditorium as well as in the space under the balcony.

In order to overcome this undesirable condition it is necessary that the rates of decay in both spaces be nearly equal (or that the rate of decay in the smaller space under the balcony be greater than the rate of decay in the main part of the auditorium). This involves a determination of the reverberation in both spaces, which in turn necessitates the assignment of coefficients of absorption to the opening which couples the two spaces. It is not possible to assign precise coefficients to these openings. The coefficients will depend, in general, upon the size and shape of the opening, the depth under the balcony, and the amounts of absorption under the balcony and in the main part of the auditorium. But if both spaces have approximately the same rates of growth, as they should for good acoustics, the "effective coefficients" will be of the order of 0.40 to 0.80—nearer the lower limit for shallow recesses which contain a relatively small amount of absorption. Similar considerations apply to the stage opening which couples the stage and the main part of the auditorium.

Many theaters, churches, memorial halls, and other auditoriums are often coupled, by means of door openings or archways, to rooms or corridors which are excessively reverberant. In such auditoriums, even though the reverberation in the audience space has been adjusted to the proper value, there will be a "feedback" of reverberation from the adjacent reverberant rooms into the main auditorium. Thus, auditors in a theater who are seated near an opening to a reverberant corridor, foyer, or anteroom will be disturbed by the excessive reverberation in the adjacent room. It is advisable in all such cases either to close the openings or to use an adequate amount of absorption in all spaces which are coupled to the audience room. In general, such anterooms, foyers, or corridors, unless they are used for speech or music rooms, should be as non-reverberant as possible.

21. REVERBERATION AT DIFFERENT FREQUENCIES

Unless otherwise specified, it is generally understood that *the time of reverberation* refers to a pure tone of 512 cycles. Although the calculation of reverberation at a single frequency, as 512 cycles, will suffice to represent the reverberation in a room at other frequencies provided the absorptive material in the room has nearly the same absorptivity at all frequencies, or provided the variation in absorptivity is known, it is obvious that such is not the case if the absorptive material has widely different and undetermined absorptivities at different frequencies. Thus, an acoutical plaster, $^1/_4$ in. thick, applied to concrete or tile, may have coefficients of absorption of 0.06 at 128 cycles, 0.36 at 512 cycles, and 0.72 at 2048 cycles. If such plaster is applied to the entire inner surface of a room the reverberation time at 128 cycles would be at least six times as long as the reverberation time at 512 cycles, and the time at 2048 cycles would be less than one-half of the 512-cycle time. If the reverberation time in such a room is 1.25 sec at 512 cycles, it will be at least 7.5 sec at 128 cycles. To describe this room as one having a time of reverberation of 1.25 sec—which is regarded as close to the optimal time for good acoustics—certainly does not describe the reverberatory properties of the room; and such a room will be highly unsatisfactory. There will be complaints of excessive reverberation, and the room will be too reverberant for the bass notes of music, and even the low-frequency components of speech will be reverberant and overemphasized. On the other hand, the higher tones and harmonics in music will be suppressed owing to over-absorption at the high frequencies. Such rooms are particularly objectionable for recording or broadcasting purposes.

It is necessary therefore to specify and calculate the reverberation times for representative frequencies throughout the entire range used in speech and music. It will be found, however, that, if calculations are made at 128, 512, and 2048 cycles, the resulting times will give a satisfactory description of the reverberatory properties of the room. In recording and broadcasting studios it is desirable to consider frequencies as high as 8000 cycles.

In general, the reverberation time at 128 cycles should be slightly longer than the time at 512, and the reverberation time above 512 should remain nearly constant. (See Fig. 10, p. 12-75, for present recommended practice.) The success or failure in the acoustical design of rooms will depend upon the selection of absorptive materials which will give the proper reverberatory characteristic throughout the entire range of frequencies used in speech and music.

22. THE MEASUREMENT OF REVERBERATION AND ABSORPTION COEFFICIENTS

The reverberation time of a room, or the total absorption of the room, for rooms in which geometric acoustics will apply, can be determined by measuring either the rate of decay of sound or the time for the decay between known intensity limits. For determining the sound-absorptive coefficients of materials it is customary to make measurements of the rate of decay of the sound in a reverberant room first when the room contains a certain area of the acoustical material to be tested and again when the material is removed from the room. In order to approximate conditions which will justify the applicability of formulas based on geometric acoustics, it is better to distribute the absorptive material on three (non-parallel) walls rather than concentrate in one area on one wall. By means of eq. (3), the value of $\bar{\alpha}$, and consequently the total absorption $\bar{\alpha}S$, can be calculated. The absorption of the acoustical material in the room is assumed to be equal to the difference between the absorption of the room with the material in it and the absorption of the room with the material removed. This is equivalent to assuming that $\bar{\alpha}$ is the arithmetical mean of all absorptive surfaces in the room, an assumption which is justifiable provided the sound in the room is kept thoroughly diffuse during the steady state and the decay. Warble tones at least 100 cycles in breadth with a warble frequency of at least 4 or 5 per sec should be used for test tones below about 500 cycles, and, unless the room be bounded by diffuse reflectors, large rotating vanes should be used for test tones of all frequencies. The test tones should be pure.

When these precaustions are taken the rate of decay will conform satisfactorily to the theoretical rate, and, if the test area is as large as about 72 sq ft in a room having a volume about 5000 to 10,000 cu ft, the difference between the rates of decay with and without the acoustical material in the room will be large enough to yield coefficients of absorption accurate to about ± 0.05 for frequencies up to 2000 cycles. At higher frequencies the absorption in the air, which may change during the time required for the completion of the test, is so large a factor that errors of the order of ± 0.10 are unavoidable unless the reverberation room is carefully air-conditioned. Even with an air-conditioned room the accuracy is not satisfactory at frequencies above 4000 cycles, because the absorption in the air is such a large factor that the difference between the rates of decay with and without the acoustical material in the room is not appreciable unless the test area is greatly increased.

The rate of decay is measured by some type of reverberation meter which in general consists of (1) a suitable source of steady or warble tones, usually a vacuum-tube oscillator, an electrical low-pass filter, a power amplifier, and an electrodynamic loudspeaker; (2) a high-quality microphone and amplifier; (3) an electrical attenuator for varying the gain of the amplifier; and (4) either a recorder which registers continuously, on a moving paper chart or on a light-sensitive medium, a graphic record of the decay, or some type of indicator, usually a relay and chronograph, by means of which the rate of decay can be determined.

The Bell Telephone Laboratories, Inc., among others, have developed a high-speed level recorder which gives a response proportional to the logarithm of the actuating current, and the instrument is so adjusted that the record gives directly, when the paper tape is moving with constant speed, the rate of decay of the sound in decibels per second. If the decay follows the exponential law, the curves will be straight lines. However, since the decay consists of several contiguous frequencies (normal modes of vibration) in close proximity to the frequency of the exciting tone, there will be interference between these several frequencies (each of which decays exponentially) so that the resultant decay curve generally will be quite irregular. Typical decay curves obtained with this instrument in the reverberation room of the Bell Telephone Laboratories are reproduced in Fig. 3. As these

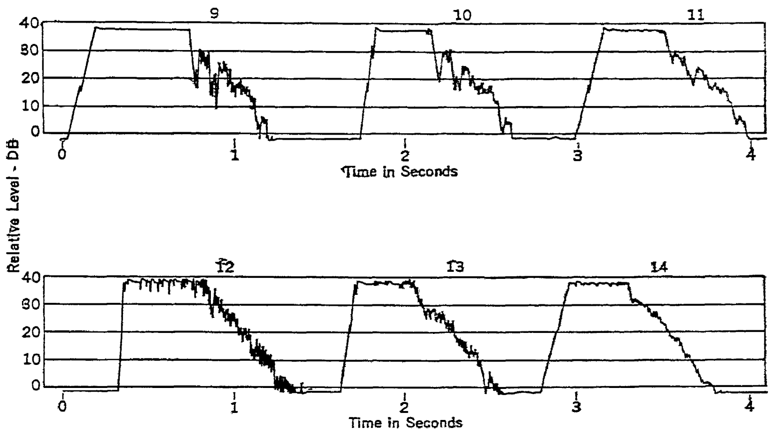


FIG. 3. Decay Curves Obtained with Bell Laboratories High-speed Level Recorder

records show, the decay is not strictly exponential but, except for minor fluctuations which can be attributed to the resonant or interfering phenomena discussed in article 20, the general trend of the decay conforms very satisfactorily to the exponential law, over a range of 40 db; and, if a straight line is fitted to the recorded curve of decay, the slope of this line will give the rate of decay with sufficient accuracy for practical purposes.

The curves labeled 9, 10, and 11 were made with a pure tone and a single microphone and with the recorder adjusted to "follow" maximal speeds of decay of 240, 120, and 60 db per sec, respectively. In 9, for example, the recorder is capable of following the actual decay much more closely than it is in 11, where only the slower variations of decay are recorded. The curves labeled 12, 13, and 14 were made at the same recorder speeds, respectively, but a warble tone was used instead of a single pure tone. The advantage of the warble tone for reverberation measurements is obvious from these decay curves.

In the chronographic type of reverberation meter, in use in many laboratories, it is customary to measure the time for a given drop in level, increasing the drop in steps of 5 or 10 db. A good meter of this type is the one developed by F. V. Hunt (*J. Acous. Soc. Am.*, Vol. 5, 127 [1933], and Vol. 8, 34 [1936]), which is almost automatic. As usual, the sound source is a warble tone oscillator. Several microphone positions throughout the room are used to insure a good average, and the reverberant sound thus detected and amplified is rectified, and rapid fluctuations are filtered out. An automatic timer turns off the source, always at the same phase of the sound wave, when the sound level in the room has reached a predetermined level, indicating the time for this decay. This is repeated 40 or more times, and the average value is used for plotting the decay curve. Figure 4 is a composite curve obtained by Hunt, showing the average decay for a warble tone of 1000 ± 200 cycles throughout a course of 80 db. By comparing the observed decay curve with the dashed straight line, it will be seen that the decay is satisfactorily exponential during the first 50 db of decay, the portion of the curve which should be used in making sound-absorption measurements. The non-linear decay from 50 to 80 db probably results from the more slowly damped tangential modes of vibration, which predominate during the latter stages of the decay.

A third method of measuring the decay rate is the oscillographic, illustrated in Fig. 2, which gives a detailed picture of the frequencies, as well as the intensity, throughout the course of the reverberation.

Reverberation measurements are useful not only for determining the coefficients of sound absorption of acoustical materials in a reverberation chamber but equally for determining the reverberatory properties of all rooms. In general, a reverberation meter should be capable of making measurements at all frequencies between about 128 and 4096 cycles and should reveal the detailed nature of the decay of the sound, especially during the first 30 to 40 db of the decay. In music rooms, recording or broadcasting rooms, and theaters, it is desirable to make measurements at frequencies as high as 8000 cycles.

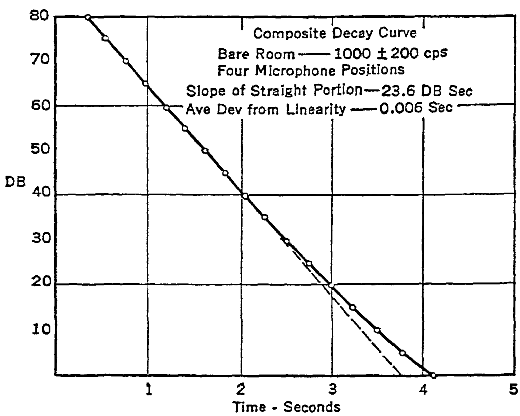


FIG. 4. Decay Curve, Based on Measurements of Average Time for a Given Drop in Level. (Hunt.)

For additional methods for measuring reverberation and absorption, consult Knudsen, *Architectural Acoustics*, Chapter VII; Sabine, *Acoustics and Architecture*, Chapter VI; Olson and Massa, *Applied Acoustics*, Chapter XII; and Beranek, *Acoustic Measurements*.

23. COEFFICIENTS OF SOUND ABSORPTION

In the following charts and tables there is given a fairly complete listing of the coefficients of sound absorption of the materials which are used in building construction, especially for acoustical purposes. Many of the same materials have been measured by different investigators, and the results are not always in good agreement. Such factors as actual differences in the samples, differences in the methods of measurement, differences in the size, shape, and location of the samples compared with the size and shape of the test rooms, the purity of the test tone used, and errors inherent in the use of reverberation methods and formulas based on geometric acoustics when wave and not geometric acoustics applies are probably responsible for this lack of agreement. Where large discrepancies have existed, certain liberties, guided by experience and the probable influence of wave acoustics, have been taken in averaging results.

Most of the materials manufactured by the leading acoustical concerns in the United States have been tested in the same laboratory by the same method; this facilitates comparison of different materials. The authority for these measurements is the Acoustical Manufacturers' Association; the tests were made at the Riverbank Laboratories. The authorities for the other measurements are listed in Table 1.

Much progress has been made in the development of acoustical materials during the past decade, and this progress will continue. Many improved and new materials will be described in subsequent issues of the *Bulletin of the Acoustical Materials Association*, which is published at frequent intervals. (These bulletins should be consulted; they can be obtained from the Acoustical Materials Association, 205 W. Monroe St., Chicago, Illinois.)

Table 1, on sound-absorption coefficients, includes considerable data concerning acoustical materials in addition to the coefficients of sound absorption, such as the *type* of the material, the nature of the *mounting*, the light reflection coefficient (usually as painted by the manufacturer), the weight per square foot of the material, and the unit size of the material. The numbers used to describe the *types* and *mountings* are those used by the Acoustical Materials Association. The legend for these numbers is given in the table.

In choosing materials for noise-reduction purposes, in offices, factories, restaurants, hospitals, etc., it is customary to rate the materials acoustically in terms of the arithmetic mean of the coefficient at 256, 512, 1024, and 2048 cycles. This average value is called the *noise-reduction factor*. It is better practice, however, to choose materials that have the absorption characteristics best adapted to the characteristics of the noise to be absorbed. It is especially important to avoid materials having very low absorption at the low frequencies for sounds which are made up predominently of low frequencies.

Table 1. Sound-absorption Coefficients

Types of Material:

I. Cast units composed of small uniform mineral particles and Portland cement.
II. Cast units having a surface composed of or resembling small uniform granules.
IV. Units having a mechanically perforated surface, backed with a sound-absorbent material.
V. Units mechanically perforated, the perforations extending into the sound-absorbent material.
VI. Units having a fissured surface.
VII. Compressed units composed of long wood fibers—surface not perforated.
VIII. Felted fiber or wood-pulp units—surface not perforated.

Types of Mounting:

1. Cemented to plaster board or other hard, rigid surface.
2. Nailed to 1 in. by 2 in. wood furring, 12 in. o.c.
3. Attached to metal supports applied to 1 in. by 2 in. wood furring.

Material	Mfr.	Thickness, in.	Type	Mounting	Coefficients 128	256	512	1024	2048	4096	Authority	Light-reflecting Coefficient	Wt., lb/sq ft	Unit Size
Acoustical Units—Tile, Wall Board, Sheets, etc.														
Absorb-a-tone..........	Luse-Stevenson	1.0	VII	1	0.11	0.20	0.58	0.91	0.81	0.77	2	0.78	2.4	36″ × 72″
Absorb-a-tone..........	Luse-Stevenson	1.0	VII	2	.15	.28	.82	.99	.87	.98	2	.78	2.4	36″ × 72″
Acousteel..............	Celotex Co.	2.5	IV	3	.25	.52	.98	.99	.81	.60	2	.73	....	12″ × 24″
Acousti-Celotex, CS-1...	Celotex Co.	0.5	V	1	.09	.15	.61	.77	.70	.64	2	.78	0.83	12″ × 12″
Acousti-Celotex, CS-1...	Celotex Co.	0.5	V	2	.14	.46	.52	.71	.72	.64	2	.78	0.83	12″ × 12″
Acousti-Celotex, C-4....	Celotex Co.	1.25	V	1	.16	.42	.99	.74	.60	.50	2	.78	1.34	12″ × 12″
Acousti-Celotex, C-4....	Celotex Co.	1.24	V	2	.25	.58	.99	.75	.58	.50	2	.78	1.34	12″ × 12″
Acoustifibre............	Nat'l Gypsum Co.	0.62	V	2	.13	.38	.72	.89	.82	.66	2	.78	0.56	12″ × 12″, 12″ × 24″
Acoustimetal...........	Nat'l Gypsum Co.	2.5	IV	3	.25	.47	.97	.99	.88	.85	2	.72	....	12″ × 24″
Acoustone F...........	U. S. Gypsum Co.	0.69	VI	1	.08	.25	.76	.84	.78	.73	2	.80	1.35	12″ × 12″
Acoustone F...........	U. S. Gypsum Co.	0.81	VI	2	.23	.50	.72	.77	.74	.69	2	.80	1.54	6″ × 12″ to 12″ × 24″
Airacoustic............	Johns-Manville	0.50	VIII	As lining ducts	.13	.41	.50	.72	.78	.72	2	....	0.80	24″ × 36″
Airacoustic............	Johns-Manville	1.0	VIII	As lining ducts	.29	.51	.70	.82	.79	.80	2	....	1.50	24″ × 36″
Auditone C...........	U. S. Gypsum Co.	0.75	V	1	.09	.28	.68	.80	.79	.81	2	.72	0.91	12″ × 12″
Auditone C...........	U. S. Gypsum Co.	0.75	V	2	.13	.49	.60	.72	.80	.81	2	.72	0.91	12″ × 12″
Auditone B...........	U. S. Gypsum Co.	1.0	V	1	.13	.33	.79	.85	.72	.70	2	.74	1.15	12″ × 12″
Auditone B...........	U. S. Gypsum Co.	1.0	V	2	.28	.48	.64	.86	.87	.75	2	.74	1.23	12″ × 12″
Balsa wood...........	....	0.44	......	2	.15	......	.19	......	.28	......	3	....	0.6	Lumber sizes

Table 1. Sound-absorption Coefficients—*Continued*

Material	Mfr.	Thickness, in.	Type	Mounting	Coefficients 128	256	512	1024	2048	4096	Authority	Light-reflecting Coefficient	Wt., lb/sq ft	Unit Size
Acoustical Units—Tile, Wall Board, Sheets, etc.														
Celotex, Standard	Celotex Co.	0.44	VIII	2	0.16	0.20	0.24	0.22	0.23	0.22	9	....	0.5	4′ boards
Cushiontone, A1	Armstrong Cork	0.50	V	1	.05	.18	.56	.76	.77	.73	2	0.78	0.79	12″ × 12″
Cushiontone, A1	Armstrong Cork	0.50	V	2	.07	.47	.55	.70	.77	.74	2	.78	0.79	12″ × 12″
Cushiontone, A3	Armstrong Cork	0.87	V	1	.09	.28	.74	.98	.78	.70	2	.78	1.17	12″ × 12″
Cushiontone, A3	Armstrong Cork	0.87	V	2	.17	.51	.73	.95	.75	.72	2	.78	1.17	12″ × 12″
Econacoustic	Nat'l Gypsum Co.	0.50	VIII	1	.05	.17	.62	.83	.77	.74	2	....	0.40	12″ × 12″
Econacoustic	Nat'l Gypsum Co.	0.50	VIII	2	.09	.32	.75	.78	.74	.78	2	....	0.40	12″ × 12″
Econacoustic	Nat'l Gypsum Co.	1.0	VIII	1	.13	.43	.78	.81	.75	.70	2	....	0.62	12″ × 12″
Fibracoustic	Johns-Manville	1.0	VIII	1	.18	.42	.81	.75	.71	.72	2	.58	0.54	12″ × 12″
Fibracoustic	Johns-Manville	1.0	VIII	2	.25	.62	.72	.72	.71	.76	2	.58	0.54	12″ × 12″
Fibretone	Johns-Manville	0.81	V	2	.18	.54	.72	.74	.71	.72	2	.71	1.17	12″ × 12″
Fir-tex Acoustical Tile (spray painted)	Fir-tex Ins. Co.	0.50	VIII	2	.11	.36	.63	.64	.67	.68	9	....	Light density	4′ boards or smaller
Fir-tex Acoustical Tile (spray painted)	Fir-tex Ins. Co.	1.0	VIII	2	.24	.51	.80	.83	.87	.89	9	....	Light density	4′ boards or smaller
Flax-li-num	....	0.75	VIII	2	.16	.24	.36	.46	.48	.46	3	....	....	Varied
Insulite	Insulite Co.	0.5	VIII	2	.22	.26	.29	.33	.37	.38	9	....	....	4′ boards
Masonite	Masonite Co.	0.44	VIII	1	.10	.21	.29	.30	.29	......	3	....	....	4′ boards
Masonite	Masonite Co.	0.44	VIII	2	.18	.25	.32	.35	.33	.31	9	....	....	4′ boards
Muffletone (standard)	Celotex Co.	1.0	II	1	.12	.30	.74	.76	.71	.67	2	....	1.80	12″ × 12″
Sanacoustic, Type KK pad	Johns-Manville	2.5	IV	3	.25	.58	.96	.97	.85	.72	2	.76	....	12″ × 24″
Simpson Acoustical Tile, Type S-2	Simpson Industries	0.62	V	2	.14	.50	.70	.83	.78	.71	2	.77	0.80	12″ × 12″
Simpson Acoustical Tile, Type S-5	Simpson Industries	1.0	V	1	.12	.31	.98	.94	.70	.64	2	.77	1.10	12″ × 12″
Simpson Acoustical Tile, Type S-5	Simpson Industries	1.0	V	2	.22	.51	.89	.98	.71	.66	2	.77	1.10	12″ × 12″
Softone	†	1.0	II	1	.07	.21	.57	.95	.89	.82	1	.85	....	12″ × 12″, 12″ × 24″
Softone	†	1.0	II	2	.10	.26	.72	.90	.75	.65	1	.85	....	12″ × 12″, 12″ × 24″
Transite Acoustical Panels	Johns-Manville	1.19	IV	2	.17	.49	.94	.90	.70	.43	2	....	....	12″ × 12″
Transite Acoustical Panels	Johns-Manville	2.19	IV	2	.29	.57	.94	.93	.70	.48	2	....	....	12″ × 12″

Acoustical Plasters and Sprayed-on Materials

Acoustipulp...........	Val-Porter Co., Los Angeles	0.50	......	‡	0.28	0.43	0.47	0.50	0.46	0.42	3	....	0.40	....
Acoustipulp...........	Val-Porter Co., Los Angeles	0.75	......	‡	.38	.51	.52	.52	.46	.42	3	....	0.60	....
Akoustolith...........	....	0.50	†	....	.17	.21	.27	.33	.38	.40	9	....	....	....
Kalite................	....	0.50	......	‡	.18	......	.30	......	.34	......	3	....	....	....
Kalite................	....	0.75	......	On metal lath	.41	......	.51	......	.59	......	3	....	....	....
Limpet (sprayed asbestos fibers)	Keasbey and Mattison Co.	0.50	......	Solid backing	.08	.27	.53	.65	.71	.59	......	....	....	....
Limpet (sprayed asbestos fibers)	Keasbey and Mattison Co.	0.75	......	Solid backing	.09	.23	.67	.90	.93	.87	......	....	....	....
Limpet (sprayed asbestos fibers)	Keasbey and Mattison Co.	0.50	......	On metal lath	.25	.78	.97	.81	.82	.85	......	....	....	....
Limpet (sprayed asbestos fibers)	Keasbey and Mattison Co.	0.75	......	On metal lath	.41	.88	.90	.88	.91	.81	......	....	....	....
Reverbolite...........	Celotex Co.	0.50	......	‡	.26	.26	.47	.57	.65	.59	2	....	1.6	....
Sabinite..............	U. S. Gypsum	0.50	......	‡	.26	.16	.32	.70	.73	.72	2	....	2.0	....

Mineral Wool Blankets, Acoustical Felts, etc.

Balsam Wool..........	....	1.0	......	1	0.12	0.25	0.52	0.81	0.67	0.53	9	....	0.26	....
Fiberglas, PF Insulation	Owens-Corning	1.0	§	2	.24	.32	.65	.77	.73	.81	2	....	0.21	....
Fiberglas, PF Insulation	Owens-Corning	1.0	§	2	.25	.41	.86	.94	.84	.81	2	....	0.50	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	1.0	‖	1	.27	.30	.57	.69	.70	......	2	....	0.25	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	1.0	‖	1	.33	.40	.76	.91	.77	.73	2	....	0.33	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	2.0	‖	1	.44	.61	.96	.93	.77	.86	2	....	0.50	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	2.0	‖	1	.55	.79	.99	.99	.91	......	2	....	1.00	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	3.0	‖	1	.55	.68	.95	.90	.79	.80	2	....	0.50	....
Fiberglas, TW-F Wool..	Fiberglas Corp.	3.0	‖	1	.69	.91	.99	.99	.91	.82	2	....	1.00	....
Hair felt.............	....	1.0	Bare	1	.12	.32	.51	.62	.60	.56	9	....	....	....
Rockwool..............	....	1.0	Bare	1	.26	.45	.61	.72	.75	......	3	....	0.90	....
Sound Isolation Blanket, MK...............	Johns-Manville	1.0	¶	1	.15	.37	.89	.98	.89	.86	2	....	1.04	....
Sound Isolation Blanket, MK...............	Johns-Manville	2.0	¶	1	.43	.64	.97	.99	.87	.90	2	....	2.33	....

† American Acoustics, Inc., Chicago, Ill.
‡ Over hard scratch and brown coats on metal lath.
§ No facing, Fiberglas blanket nailed to 1 in. by 2 in. furring, 12 in. o.c.
‖ Perforated metal facing, 26 gage, 0.076-in. holes, 0.176-in. o.c.
¶ Muslin covered, blanket on floor.

Table 1. Sound-absorption Coefficients—*Continued*

Description	Thickness, in.	Coefficients 128	256	512	1024	2048	4096	Authority
Set Materials for Use in Motion-picture Studios								
Cast plaster, applied to burlap, one coat thin shellac	0.38–0.5	0.10	0.09	0.05	0.07	0.09	0.05	3
Cast stone, similar to above, except irregular surface	....	.24	.23	.10	.15	.15	.09	3
Celotex, papered with crepe paper	.44	.17	.14	.11	.11	.12	.11	3
Celotex, as above, with one coat water paint	.44	.17	.16	.11	.11	.13	.09	3
Masonite, papered with crepe paper	.44	.18	.17	.11	.10	.12	.11	3
Masonite, as above, with one coat studio flat paint	.44	.16	.17	.11	.09	.07	.07	3
Veneered flats, papered with crepe paper	....	.12	.11	.06	.08	.09	.12	3
Veneered flats, as above, and hard wallpaper over crepe paper	....	.10	.10	.06	.07	.07	.07	3
Zonolite, ** brushed over burlap on chicken wire	.06– .12	.38	.37	.25	.23	.24	.22	3
Hangings, Floor Coverings, and Miscellaneous Materials								
Carpet, lined	....	0.10 ††	....	0.25	....	0.40 ††	....	5
Carpet, unlined	....	.08 ††	....	.15	....	.25 ††	....	5
Carpet, Amritza, on concrete	.44	.09	.06	.24	.24	.24	.11	6
Carpet, Cardinal Batala, on concrete	.44	.12	.10	.28	.42	.21	.33	6
Carpet, pile, on concrete	.38	.09	.08	.21	.26	.27	.37	6
Carpet, pile, on 1/8-in. felt	.38	.11	.14	.37	.43	.27	.25	6
Carpet, rubber, on concrete	.19	.04	.04	.08	.12	.03	.10	6
Cork flooring slabs, glued down	.75	.08	.02	.08	.19	.21	.22	6
Cork flooring, like above, waxed and polished	.75	.04	.03	.05	.11	.07	.02	6
Cotton fabric, 14 oz per sq yd, draped to half its area	....	.07	.31	.49	.81	.66	.54	4
Cotton fabric, like above, draped to seven-eighths its area	....	.03	.12	.15	.27	.37	.42	4
Draperies, velour, 18 oz per sq yd	....	.05	.12	.35	.45	.38	.36	4
Draperies, like above, draped to one half area	....	.14	.35	.55	.72	.70	.65	3
Linoleum, on concrete floor	....	.02 ††	....	.03	....	.04 ††	....	3
Openings, balcony	....	....		.25 to .80		....	....	9
Openings, stage, depending upon stage furnishings	....	....		.25 to .40		....	....	7
Oregon pine flooring	.75	.09	....	.08	....	.10	....	3
Ozite, 0.39 lb per sq ft	.5	.06	.13	.20	.42	.47	.47	4
Rug, Axminster	....	.11	.14	.20	.33	.52	.82	8
Ventilators, 50 per cent open	....	.30 ††	....	.50	....	.50 ††	....	7

Hard Plasters, Masonry, Wood, and Other Standard Building Materials

Brick wall, unpainted	18.0	0.02	0.02	0.03	0.04	0.05	0.07	5
Brick wall, painted	18.0	.01	.01	.02	.02	.02	.02	5
Glass	...	.03 ††	...	.03	...	.02 ††	...	5
Interior stucco, smooth finish, on tile	0.5	.03 ††	...	.04	...	.04 ††	...	3
Marble	...	.01 ††	...	.01	...	.01 ††	...	7
Plaster, gypsum, on hollow tile	...	.01	.01	.02	.03	.04	.05	5
Plaster, gypsum, scratch and brown coats on metal lath, on wood studs	...	.02	.03	.04	.06	.06	.03	4
Plaster, lime, sand finish, on metal lath	0.75	.04	.05	.06	.08	.04	.06	3
Poured concrete, unpainted	...	.01	.01	.02	.02	.02	.03	3
Poured concrete, painted and varnished	...	.01	.01	.01	.02	.02	.02	3
Water, as in swimming pool	...	.01	.01	.01	.01	.02	.02	3
Wood sheathing, pine	0.75	.10	.11	.10	.08	.08	.11	5
Wood veneer, on 2 in. by 3 in. wood studs, 16 in. o.c.	0.44	.11	...	.12	...	.10	...	3

Audience, Individual Persons, Chairs, and Other Objects

Audience, mixed, seated in theater chairs, single padding on back	...	...	3.5	4.1	4.9	4.2	...	1
Audience, mixed, seated in church pews	...	...	2.7	3.3	3.8	3.6	...	1
Chair, American loge, full upholstered in mohair	...	...	...	4.5	...	...	...	1
Chair, box spring, pantasote seat and back, plywood on rear; seats up	...	...	1.4	1.6	1.3	0.71	...	7
Chair, plywood seat, plywood back; seats up	...	...	0.19	0.24	0.39	0.38	...	7
Chair, spring edge mohair seat and back, plywood panel on rear; seats down	...	...	3.1	3.0	3.3	3.5	...	7
Chair, like above, with thick, completely covered seat and back; seat up	...	...	3.3	3.5	3.7	3.8	...	7
Chair, theater, heavily upholstered	...	...	3.4	3.0	3.3	3.6	...	1
Person, adult	...	1.8	...	4.2	...	5.5	...	3
Person, adult, seated in American loge chair	...	...	...	5.5	...	...	...	1
Person, child, high school	...	1.6	...	3.8	...	5.0	...	3
Person, child, junior high school	...	1.5	...	3.5	...	4.6	...	3
Person, child, grammar school	...	1.3	...	2.8	...	3.8	...	3
Person, man, with coat, seated	...	2.3	3.2	4.8	6.2	7.6	...	1
Person, woman, without coat, seated	...	0.7	1.3	2.3	3.6	4.6	...	1
Person, woman, with coat, seated	...	1.3	2.4	4.0	5.8	6.7	...	1

** Small granules of exfoliated vermiculite bonded with a plastic, tacky binder.

†† These coefficients are estimates made by the compiler.

Authority: (1) Bureau of Standards, (2) Acoustical Materials Association, (3) V. O. Knudsen, (4) P. E. Sabine, (5) W. C. Sabine, (6) Building Research Station, (7) F. R. Watson, (8) Wente and Bedell, (9) average.

The table contains absorption data on a number of mineral wool blankets, made up of wool of different densities, and of thicknesses varying from 1 to 3 in. It will be seen that by suitable choice of density and thickness it is possible to obtain a wide variety of absorption characteristics.

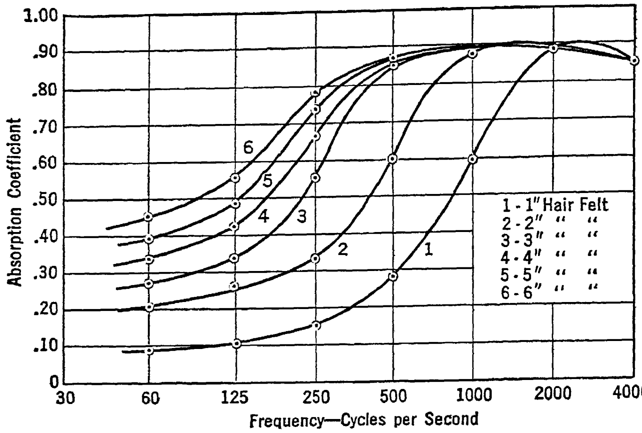


FIG. 5. Absorption of Different Thicknesses of Hair Felt. (Wente and Bedell.)

In Fig. 5 are shown the results of sound-absorption measurements of Wente and Bedell on hair felt of different thicknesses. The principal effect of increasing the thickness of a porous material is to increase the absorption at the low frequencies.

In Fig. 6 are shown the results of measurements at 512 cycles on different thicknesses of hair felt obtained in four different laboratories. These results show not only the effect of thickness but also the order of agreement of measurements made at different laboratories.

The manner of mounting acoustical materials influences their absorptivity. Most fibrous and porous materials increase in absorptivity, especially at frequencies below 500 cycles, as the thickness of the air space behind the material increases. Thus, fiber board

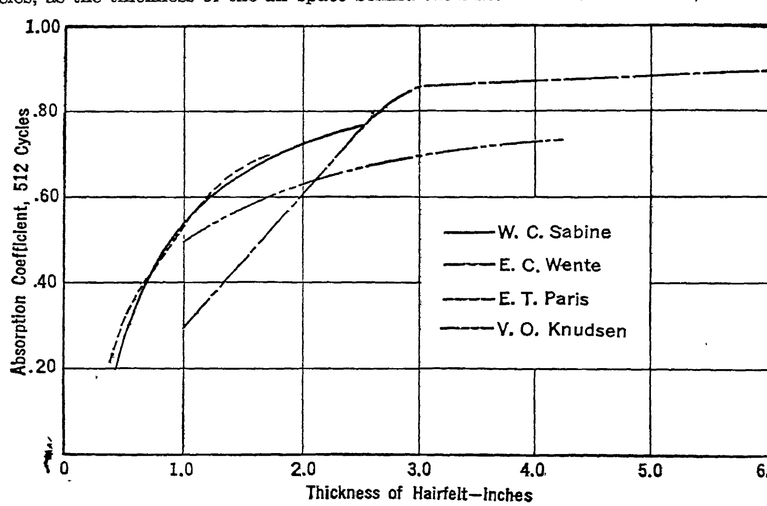


FIG. 6. Absorption of Hair Felt as Determined in Different Laboratories

and tiles, acoustical plaster, and similar materials, are more absorptive at low frequencies when furred out from a dense, rigid wall than when applied directly to the wall.

The absorptivities of such materials as acoustical plaster are dependent upon the composition as well as upon the manner of applying and drying. If too much binder material is used the plaster is not sufficiently porous; if an insufficient amount of binder is used the plaster does not set hard. Likewise, if the undercoats of plaster are too wet (or "green"), the binder material forms an impenetrable film at the surface, whereas if the undercoats are too dry the binder material is absorbed by the undercoats, and the plaster will crumble. In order to obtain good results with acoustical plaster it must be applied by competent plasterers in strict conformity with the specifications of responsible manufacturers.

The absorptivity of acoustical plaster or fiber board may be ruined by decoration with oil or water paint, varnish, distemper, or other materials which will close the surface pores. Viscous or heavy paints which *bridge over* or close the surface pores must be avoided. Such materials as thin aniline dyes, gasoline or kerosene stains, thin lacquer sprays, or dry paint dusted on with a pounce bag are satisfactory means of decoration without impairing the absorptive properties of plaster or fiber board. Certain commercial materials containing large holes, or materials covered with a perforated facing, may be decorated with lead or oil or any other kind of paint, provided the paint does not bridge over the holes.

In order to facilitate convenient use of the table the materials have been grouped as follows: acoustical tiles, boards, and sheets; acoustical plasters and sprayed-on plastic materials; mineral wool, Fiberglas blankets, and acoustical felts; set materials for use in motion-picture studios; hangings, floor coverings, and miscellaneous materials; hard plasters, masonry, wood, and other standard building materials; and audience, individual persons, chairs, and other objects. The tabulation in each group of materials is arranged in alphabetical order.

24. PRACTICAL CONSIDERATIONS OF SOUND-ABSORPTIVE MATERIALS

In making a choice of absorptive materials a number of other factors must be considered besides the coefficients of sound absorption. Good acoustics is only one of many qualities which should be secured in every building. Thus, besides absorption coefficients, it is necessary to consider such factors as the following: structural strength; decorative possibilities; adaptability to the surface available for, or requiring, absorptive treatment; maintenance; sanitation; ease of application; fire hazard; absorption of water; attraction for vermin; "fool-proofness"; durability; and cost. Each room requires a certain amount of sound absorption; certain surfaces in some rooms require highly absorptive treatment. These two conditions usually limit the choice of absorbents to materials having coefficients within certain specified limits. As a rule, however, many materials having coefficients within these limits will be available. This allows considerable freedom in the selection of materials which will be satisfactory not only acoustically but also for the other requirements. For a detailed discussion of the choice of sound-absorptive materials consult any standard textbook on architectural acoustics (as Knudsen and Harris, *Acoustical Designing in Architecture*).

For bibliography, see p. 12-76.

SOUND INSULATION

By Vern O. Knudsen

25. NOISE MEASUREMENTS

During recent years, acoustical engineers and civic authorities have become increasingly aware of the problems associated with the measurement and abatement of noise. One of the prime requirements for good acoustics in every room is absence of noise, i.e., *unwanted sound*. In the design of theaters, music rooms, churches, schools, office and industrial buildings, hotels, apartment houses, and studios for the recording or broadcasting of sound, it is necessary that the designing architect or engineer know (1) the amount and kind of the noise against which he is to provide insulation, and (2) the amount of noise which can be tolerated in different types of buildings. The difference between the magnitudes of (1) and (2) gives the amount of sound insulation which should be provided in the building. As discussed in article 5 of this section the magnitudes of noises vary over a wide range.

The magnitude and character of steady sounds can be represented by plotting the intensity level (per cycle or for a specified band width) as a function of the sound frequency. In Fig. 1, the intensity level per octave, i.e., the intensity levels as measured in octave bands, for a number of ordinary sounds are plotted for the important audio-frequency range of 50 to 10,000 cycles. Traffic noise will be seen to have its predominant intensity at low frequencies; the noise from typewriters, on the other hand, is greatest at high frequencies. Such characteristics of noise should be considered in the problems of sound insulation and noise reduction in building design.

Most measurements of noise, of interest in building design, have been made with commercial sound-level meters which measure the overall sound level in decibels rather than the intensity level as a function of the frequency. The latter should be used whenever available, but the sound level corresponds roughly with the *sensation* of the sound and provides a convenient numerical scale for comparing the levels of different sounds. Thus, Fig. 2 gives the average sound levels, in decibels, as measured with a sound-level meter, for a large number of locations in or near buildings.

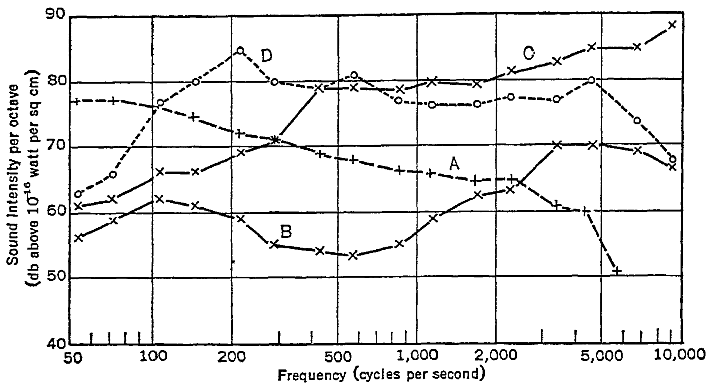


FIG. 1. Sound Spectra of Some Typical Noises. (Fleming and Allen.) (British Crown Copyright reserved. Reproduced by permission of the Controller of His Britannic Majesty's Stationery Office.)

Examples of Noise Analysis

A—Traffic noise. Average of miscellaneous vehicles passing at about 20 ft.
B—Typists' room. Two typewriters in operation.
C—Woodwork shop. 14-in. circular saw.
D—Woodwork shop. Planing machine.

If a sound wave completely modulates the pressure of the air at sea level, i.e., if the instantaneous pressure varies from 0 to 2 atmospheres, the intensity level would be 194 db. This represents an upper possible limit for the intensity of sounds, a limit which is not far above that actually attained with the Victory Siren (a large air stream modulated with a "chopper"), which was used in New York City and other American cities as an air-raid alarm during World War II. Most of the other entries in Fig. 2 are self-explanatory. The standard deviations are given for a number of measurements; these were made by Bell Telephone Laboratories (D. F. Seacord, *J. Acous. Soc. Am.*, Vol. 12, 183) at more than 600 locations in four different cities. The measurements in the private hospital room (made by the author with a continuous recorder) revealed sound levels of 50 to 58 db during 80 per cent of the time from 5:00 P.M. to 7:00 P.M.; at 9:00 P.M. the level had dropped 2 or 3 db. In the corridors of this same hospital, levels of 65 to 75 db were common; there were peak levels of 78 db from the closing of elevator doors 55 ft away from the sound-level recorder, and 90 db from coughing 40 ft away.

26. ACCEPTABLE NOISE LEVELS IN DIFFERENT BUILDINGS

Although it is not customary for building codes to specify the allowable noise in different types of buildings, and opinions differ considerably as to tolerable noise levels, the following table gives approximate loudness levels which will be, in general, highly satisfactory:

	Decibels
Radio, recording and television studios	25 to 30
Hospitals	35 to 40
Music rooms	30 to 35
Apartments, hotels, and homes	35 to 45
Theaters, churches, auditoriums, classrooms, and libraries	30 to 40
Private offices and conference rooms	35 to 45
Large public offices, banking rooms, stores, etc.	45 to 55
Restaurants	50 to 55
Factories	45 to 80

Special conditions or circumstances, such as past experience, other near-by noises, and costs, may alter the acceptable noise levels, but the levels given in the table are recommended for purposes of building design. As will be seen by comparing these values with those given in Fig. 2 the acceptable values given in the table are somewhat lower than

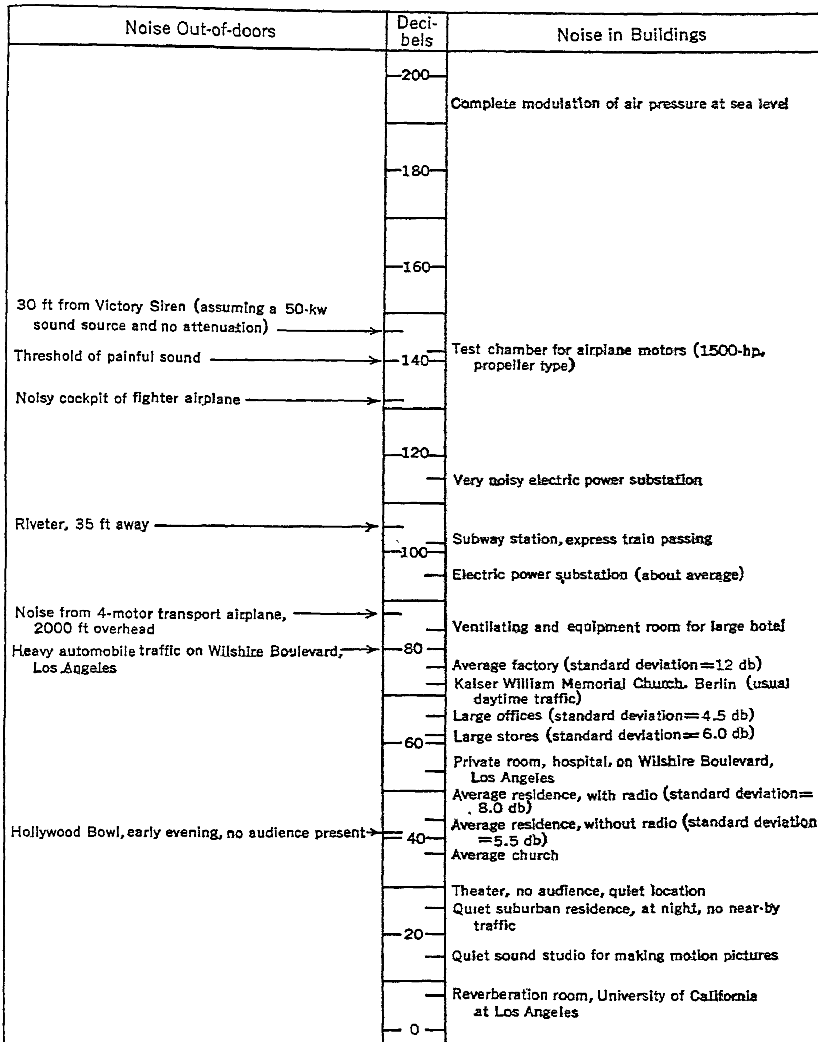


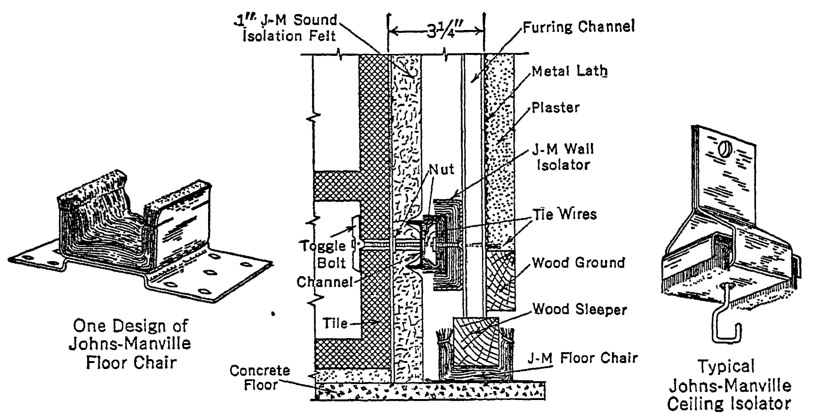
FIG. 2. Sound Levels in or Near Buildings

those that prevail in existing buildings. The average loudness level in the private hospital room, for example, is 54 db, which is some 15 to 20 db greater than the level proposed in the table. Since the outside traffic noise adjacent to this hospital has an average level, when traffic is heavy, of 80 db, and since the average tolerable level for a hospital is 37 db, the building should be designed for sound insulation in such a manner as will provide an overall noise reduction 80–37 db, i.e., 43 db. Part of the required noise reduction can be accomplished by a proper "setback" of the building and, in some instances, by dense planting of evergreen shrubs and trees between the street and the building; another part, by the use of sound-absorptive materials within the building; but most of the reduction can be accomplished only by proper sound insulation of the building itself.

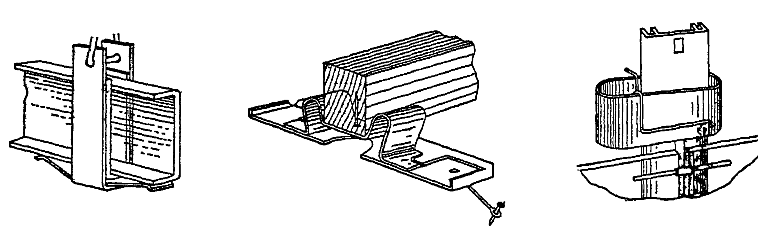
27. FUNDAMENTAL PRINCIPLES OF SOUND INSULATION

Nearly every building is subject to the annoyance of noises which have their origin in adjacent rooms or outside. It is possible to design buildings in such a manner as to exclude effectively both these types of annoying noises. The principal means whereby such noises enter a building are the following:

1. By means of openings, as windows, cracks around doors, ventilating ducts, or any other openings that will admit a free flow of air.
2. By means of refraction or transmission through partitions. This is analogous to the refraction or transmission of light from air to water, or between any other two dissimilar media.
3. By means of the conduction of sound through solids. For example, "impact sounds," such as footfalls, hammering on walls or floors, or the moving of furniture on hardwood floors, are conducted through the dense and rigid structural members of a building.
4. By means of the diaphragm action of walls which communicate sound from one side of a partition to the other side.



Typical Johns-Manville Wall Isolating Treatment



As Used by United States Gypsum Co.

Fig. 3. Flexible Cushions, Supports and Connectors

The refraction or transmission of sound from one medium to another, as from air to plaster or stone, is an almost negligible factor in building construction—not more than about one-millionth of the intensity of the incident wave in air can penetrate a material like brick or stone.

The transmission of sound through openings, on the other hand, is often the means by which sound is most readily transmitted from one portion of a building to another. This means of transmission often limits the amount of sound insulation which can be obtained in buildings, especially in hotels and apartments where the insulation is determined by such unavoidable openings as may be incidental to the use of windows and doors. Under such circumstances it would be futile to provide a relatively high insulation through the separating walls or partitions. Even very small openings, such as cracks around doors or around imperfectly fitting windows, are effective in transmitting a considerable amount of sound.

The transmission of sound through ventilating ducts often becomes a troublesome problem. There are three types of sound-transmission which must be controlled: (1) the noise from the fans, motors, and other air-conditioning equipment which is transmitted through the ducts and into the room; (2) the noise from an adjacent room which is transmitted from opening to opening, often through a short and highly conductive section of duct; and (3) the noise from adjacent rooms or outside which may be transmitted through the walls of the duct and thence through the ducts and into the room. The noise from the ventilating equipment room can be reduced suitably by (1) the selection of slow-speed, quietly operating equipment; (2) treating the walls and ceiling of the equipment room with highly absorptive material; and (3) introducing acoustical attenuation within the ducts, which can be accomplished by using very long ducts of small cross-sectional area and by lining the ducts with highly absorptive material, or by introducing other acoustical filters in the ducts.

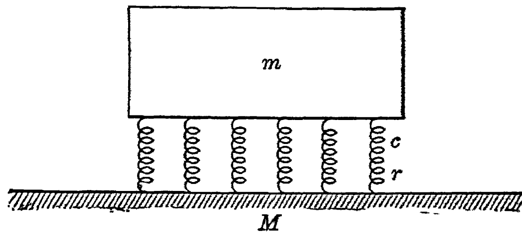


FIG. 4. Effective Method for Insulation of Vibration

Solid-borne sound travels through the structural members of a building with but very little attenuation. The compressional wave in a solid is often communicated to large surfaces, as the walls or floor of a room, and these large surfaces are made to vibrate like the sounding board of a piano. In this way a large portion of the solid-borne sound may be radiated into a room. Many solid-borne sounds can be controlled by the carpeting of floors; by the wrapping of pipes—especially where they touch the frame of the building—with flexible, porous materials, as hair felt; or by the proper mounting of machinery on flexible or elastic supports, so that the natural frequency of the machinery mounted on its flexible support will be low in comparison with the frequencies which are to be insulated. One of the most effective methods of eliminating these solid-borne sounds is to introduce discontinuities in the paths of the conducted sounds. These discontinuities should consist of materials which differ largely in elasticity and density from the solid structure of the building. For example, it is possible to suspend the ceiling by means of flexible supports; it is possible to build up inner walls in a room which are fastened to the monolithic frame by means of flexible ties; and it is possible to float the floor of the room upon flexible pads of cork or felt or other elastic material. In Fig. 3 are shown two types of flexible cushions, supports, and connectors which are effective for insulating solid-borne vibrations in buildings. Figure 3 also shows some constructional details for insulating solid-borne sound in buildings.

Figure 4 shows a simple method of insulating any object, as a part of a building or a piece of equipment, from earth, building, or machinery vibration. The problem consists of insulating a mass m from another object of mass M by means of a flexible support which has certain elastic and damping properties. Figure 5 is the electric circuit equivalent

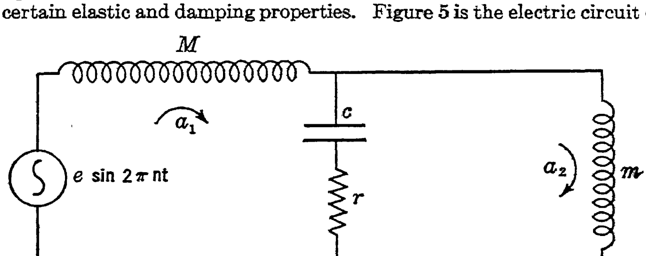


FIG. 5. Electric Circuit Equivalent of Fig. 4

of Fig. 4. This circuit implies that, when M is set into forced periodic vibration, these vibrations are communicated to m principally by means of the elastic coupling between M and m, although the internal damping or resistance r of the system also contributes to the coupling. If a_1 and a_2 represent the amplitudes of vibration set up in M and m, respectively, n the frequency, and c the compliance, then

$$\frac{a_2}{a_1} = \sqrt{\frac{r^2 + (1/4\pi^2n^2c^2)}{r^2 + [2\pi mn - (1/2\pi nc)]^2}} \tag{1}$$

This equation has been tested experimentally for both supported and suspended systems and is in good agreement with the observed results. The equation is useful for calculating the insulation value of different types of flexible supports. For values of n which are small compared with the natural or free vibration of m upon its elastic support, a_2/a_1 will be equal to unity; that is, m and M vibrate with the same amplitude. At the resonant frequency, a_2/a_1 is greater than unity, or the *insulating* support actually amplifies the motion imparted to m. However, at frequencies greater than $\frac{1}{\pi}\sqrt{mc}$ the value of a_2/a_1 becomes less than unity, and it approaches the value $r/2\pi nm$ at frequencies which are high compared with the natural or resonant frequency. In general, both m and c should be as large as possible if m is to be well insulated from the vibrations in M; that is, the support should be very elastic and loaded as heavily as possible (see also article 30).

In Table 1 are given the values of the compliance c and the resistance r of a number of materials which are used for the insulation of vibration. Beside these materials there are a number of patented devices, similar to those shown in Fig. 3, which are very effective. (See *Sound Transmission in Buildings*, by Fitzmaurice and Allen, His Majesty's Stationery Office, London [1939], and *Modern Theory and Practice in Building Design*, by Fleming and Allen, The Institution, London [1945].)

Table 1. Compliance and Resistance Data for Typical Specimens of Flexible Materials

The compliance and resistance given in the table are for specimens 1 in. thick and 1 sq cm in cross-section.

Material	Description of Material	Approximate Upper Safe Loading, lb per sq in.	Compliance, c, cm per dyne	Resistance, r, absolute units
Corkboard	1.10 lb per board ft	12	0.25×10^{-6}	0.15×10^5
Corkboard	0.70 lb per board ft	8	0.50×10^{-6}	0.25×10^5
Flax-li-num	1.35 lb per board ft	4–6	0.60×10^{-6}	0.50×10^5
Celotex	Insulating board	12	0.18×10^{-6}	
Insulite	Insulating board	15	0.16×10^{-6}	
Masonite	Insulating board	15	0.12×10^{-6}	
Sponge rubber	25 lb per cu ft	1–3	3.0×10^{-6}	
Soft India rubber	55 lb per cu ft	3–6	1.2×10^{-6}	
Hair felt	10 lb per cu ft	1–2	1.5×10^{-6}	

In the choice of materials for the insulation of vibration or solid-borne sound, it is necessary to give consideration to the safe amount of loading the material will withstand without breaking down or without being compressed to the extent that its compliance is reduced beyond required limits. It also is important to select a material that will have a long life and that will not continue to compress or settle under the load which it supports. For example, if ordinary insulation cork is loaded as much as 20 or 30 lb per sq in., the material will continue to compress indefinitely, and at the same time will become less and less compliant, until ultimately it not only loses its insulation value but also allows an amount of settling which cannot be tolerated. For example, a specimen of 1-in. insulation cork (0.70 lb per board ft), under a load of 20 lb per sq in., settled 0.04 in. during the first 24 hours, 0.02 in. during the next 24 hours, and 0.11 in. during the first 5 months it was under compression. The same specimens, under a load of 10 lb per sq in., settled only 0.01 in. during the first 24 hours, 0.005 in. during the next 24 hours, and only 0.03 in. during the first 5 months. In general, the most satisfactory material will be one that has a high compliance and very little tendency to settle under the influence of the load and that tends to return to its initial condition when the load is removed. Hair felt, cork, and rubber seem to be the best available materials that meet these requirements, although all these materials continue to settle, and become less and less compliant, as they become older. Flexible steel supports and clips, such as those shown in Fig. 3, do not have these defects and are proving to be very satisfactory not only for the insulation of walls, floors, and ceilings but also for insulating all sorts of equipment from the floor or the rigid frame of the building.

INSULATION OF SOUND BY POROUS MATERIALS. The insulation of sound by porous materials is accomplished principally by viscous losses within the capillary pores within the material and by the vibration of the component parts of the material. Figure 6 shows the transmission coefficients at different frequencies for one, two, three, and four layers of hair felt having a density of 12 lb per cu ft. These results show that, approximately, the logarithm of the energy reduction, or the reduction in decibels, of sound transmitted through porous materials is proportional to the thickness of the material. The results also show that porous materials, if used by themselves, do not provide a very high degree of sound insulation unless the insulating blanket or partition is very thick. Thus, at 700 cycles, the coefficient of transmission for four layers of hair felt—each layer is 0.58 in. thick—is about 0.01; that is, a sound wave of 700 cycles would be attenuated only 20 db in passing through 2.32 in. of hair felt. However, when such materials are used properly in conjunction with rigid partitions, they may contribute a considerable amount to the total insulation supplied by a wall structure. One of the most effective ways in which such materials may be used for the insulation of sound is by suspending or supporting the porous blanket in an air space between two rigid partitions.

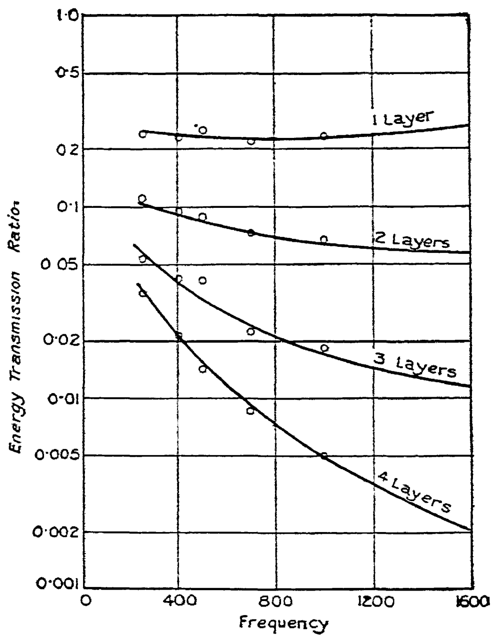


Fig. 6. Transmission Coefficient for One, Two, Three, and Four Layers of Hair Felt. (Davis and Littler.)

INSULATION OF SOUND BY RIGID PARTITIONS. Sound is transmitted through rigid partitions principally by the forced vibration of the wall; that is, the entire partition is forced into vibration by the pressure variations of the incident sound wave. The transmission coefficient τ for a heavy partition, as masonry or concrete, for normally incident sound, is given, approximately, by

$$\frac{1}{\tau} = 1 + \frac{1}{4}\left(\frac{\rho_1 c_1}{\rho c}\right)^2 \sin^2 k_1 l \quad (2)$$

where ρ_1 and c_1 are the density and sound velocity in the partition, ρ and c the corresponding density and velocity in the air, $k_1 = 2\pi/\lambda_1$, where λ_1 is the wavelength of the sound in the partition and l is the thickness of the panel. The transmission loss in decibels ($10 \log_{10} 1/\tau$) for a 9-in. brick wall, as calculated by eq. (2), is about 50 db at 100 cycles and rises almost uniformly with frequency to 78 db at 4000 cycles, and then drops to a minimum at 8000 cycles (which is zero according to eq. [2] but actually is of the order of 50 db because of dissipation within the partition). At frequencies below about 100 cycles, the resonant properties of the partition are of considerable influence; the transmission loss (T.L.) at these low frequencies may be much less than that calculated by eq. (2). The observed T.L. for heavy masonry partitions, even at 100 to 4000 cycles, is some 10 to 15 db less than that calculated from eq. (2), but the equation is useful for predicting the influence on T.L. of such factors as the wavelength of the sound and the mass and thickness of the partition.

For most rigid walls in buildings, as concrete, brick, clay or gypsum tile, wood or metal studs plastered on one or both sides, and even glass or metal panels, eq. (2) can be further simplified so that, approximately, for frequencies of 100 to 4000 cycles,

$$\frac{1}{\tau} = \left(\frac{1}{4}\right)\left(\frac{k^2 m^2}{\rho^2}\right) \quad (3)$$

where $k = 2\pi/\lambda$ (λ is the wavelength of the sound in air); and m is the effective mass per unit area of the wall, which is of the order of 0.2 to 1 times the actual mass per unit area. The mass reaction of rigid walls is thus the dominant factor affecting sound transmission; the sound is transmitted largely by the diaphragmlike vibration of the wall, the wall responding to the alternating force of the impinging sound wave much as a rigid piston would.

In thin flexible panels, the stiffness, the internal damping, the size of the panel, and the manner in which it is clamped around the edges all contribute to the total amount of vibration which will be imparted to the partition, and therefore all these factors contribute to the insulation value of such walls. However, these factors are effective principally at low frequencies; the mass is the predominant factor throughout most of the audio-frequency range in determining the insulation value of nearly all rigid panels and partitions encountered in practice. In Fig. 7 is given a summary of the measured insulation value of many rigid panels having weights varying from 2 lb per sq ft up to 100 lb per sq ft. There is a nearly linear relation between the average T.L. and the logarithm of the weight per square foot of the partition. (The average T.L. is the arithmetical mean of the measured transmission losses at 128, 256, 512, 1024, and 2048 cycles.)

Whereas the insulation value of porous materials is proportional approximately to the thickness of the material, the insulation value of a rigid material increases only as the logarithm of the thickness. Because of the slow increase in insulation with increased mass or thickness of a rigid partition, it is not always feasible to secure a high insulation by merely increasing the thickness of the wall. Thus, it would be necessary to increase the thickness of a concrete wall to nearly 4 ft in order to give the wall an insulation of 60 db.

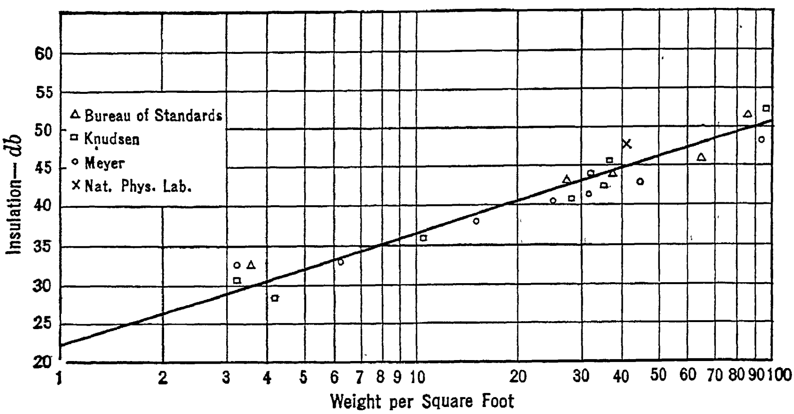


FIG. 7. Transmission Loss in db for Rigid Single Partitions

When walls of high insulation are required, it is more feasible and economical to employ special structures which combine the two principles of sound insulation just described: namely, absorption losses in porous-flexible materials, and inertial losses in massive partitions. Thus, two or three rigid and relatively thin partitions separated from each other by felts or blankets can easily be composed in such a manner as to give an insulation of 60 or even 70 db. The insulation value of many special forms of construction employing these and other principles will be found in the tables in the next section. (See also references given on p. 12-75.)

28. COEFFICIENTS OF SOUND TRANSMISSION

The coefficient of sound transmission of a panel is the ratio of the transmitted to the incident sound energy. One of the most satisfactory methods for measuring the coefficients of different materials is to make measurements of the sound intensity on both sides of a test panel placed between two rooms which are so constructed that no sound is transmitted from one room to the other except through the test panel. Thus, suppose the average intensity near the panel in the source room to be I_1 and the average intensity in the test room to be I_2. Then the rate of flow of sound energy against the test panel in the source room will be I_1A, where A is the area of the test panel. The rate at which energy will be transmitted through the panel into the test room will be $I_1A\tau$, where τ is the transmission coefficient for the panel. The product $I_1A\tau$ is then the rate of emission of sound energy in the test room. Hence, when equilibrium is established, this rate of emission of sound energy will be equal to the rate of absorption of sound in the test room, which is equal to I_2a, where a_2 is the total absorption of the test room. Hence,

$$\tau = \frac{I_2a_2}{I_1A} \tag{4}$$

The coefficient of transmission thus involves not only the intensities on the two sides of the panel but also the area of the panel and the total amount of absorption in the test room.

Table 2. Coefficients of Sound Transmission

Description of Panel (Bold-face type signifies that the panel has practical merit)	Weight, lb per sq ft	Reduction Factors in Decibels					Authority	Probable Average T.L., db	Probable Average Value of τ
		128	256	512	1024	2048			
Porous-flexible Materials and Fiber Boards									
Celotex, Standard, 0.5 in.	0.66	......	22.4	17.3	23.4	27.4	1 *	20	0.010
Insulite, 0.5 in.	0.75	......	22.2	20.2	24.1	20.9	1 *	19	0.013
Hair felt, 1 in.	0.75	4.9	4.6	6.0	7.1	6.7	3	......	....
Hair felt, 4 in.	....	7.5	12.5	15.3	19.7	19.4	3	......	....
Rock Wool blanket, 0.5 in. covered on both sides with heavy brown paper	....	15.5	......	17.8	......	18.4	2	16	0.025
Upson Blue Stripe Insulation	....	14.0	15.1	16.0	18.5	21.1	2	16	0.025
Thin Rigid Materials									
Aluminum, 0.025 in.	0.35	......	17.9	13.2	17.7	23.2	1 *	16	0.025
Duralumin, 0.020 in.	0.33	......	14.1	12.5	17.6	22.5	1 *	15	0.032
Iron, 0.03 in. galvanized	1.2	......	25.3	20.5	28.8	35.0	1 *	25	0.0032
Lead, 0.062 in.	3.9	......	31.8	33.2	32.0	32.1	1 *	30	0.0010
Plywood, 0.25 in., three-ply	0.73	......	21.0	20.7	25.5	26.0	1 *	21	0.0080
Mahogany, 1.85 in.	4.9	......	26.0	27.0	36.0	......	4 †	28	0.0016
Plaster board, 0.5 in.	....	27.0	......	28.0	......	33.0	2	28	0.0016
Doors and Windows									
	Doors								
Birch veneer, light, four panel	....	13.0	16.1	20.4	22.8	22.0	3	22	0.0063
"Cold-storage" door, double wall, 4 in.	....	16.4	20.8	27.1	29.4	28.9	3	29	0.0013
Oak, solid, 1175 in., with cracks as ordinarily hung	....	11.5	15.1	20.4	22.0	16.2	3	20	0.01
Oak, like above, well seasoned and air tight	....	15.1	18.2	22.8	25.7	25.2	3	25	0.0032
Steel, solid, 0.25 in.	....	25.1	26.7	31.1	36.4	31.5	3	35	0.00032
	Windows								
Glass, plate, 0.25 in.	3.5	......	32.6	30.9	33.5	34.2	1 *	30	0.0010
Glass, plate, 0.25 in., four panes	....	23.2	20.8	26.4	27.5	22.8	3	29	0.0013
Glass, plate, 0.25 in., double glazed, 16 in. separation	....	......	......	43.0 ‡	......	......	3	48	0.000016
Rigid Partitions (Tile, Brick, Concrete, etc.)									
Brick panel, 8 in.; plastered both sides	97.0	......	47.7	49.4	57.0	59.2	1 *	49	0.000013
Tiles, hollow-clay partition, three cells, 4 in. by 12 in. by 12 in., plastered both sides	29.0	......	41.1	40.0	41.5	49.9	1 *	40	0.00010
Tile, hollow clay, 4 in., unplastered	17.0	24.5	24.1	26.1	35.5	29.8	3	35	0.00032
Tile, like above, 0.5 in. plaster	22.0	25.1	24.3	26.9	38.2	33.9	3	38	0.00016
Tile, hollow gypsum, 3 in., unplastered	11.1	19.2	18.7	20.8	28.5	30.0	3	31	0.00080
Brick 4.5 in., 0.5 in. fiber board stuck on each face, then 0.6 plaster on each face	46.8	......	44.	45.	51.	73.	6	48	0.000016
Brick, 0.25 in. Masonite on lath on one side, 11 in.	88.0	36.5	37.0	48.0	58.5	61.0	5 §	48	0.000016

Table 2. Coefficients of Sound Transmission—*Continued*

Description of Panel (Bold-face type signifies that the panel has practical merit)	Weight, lb per sq ft	Reduction Factors in Decibels					Authority	Probable Average T.L., db	Probable Average Value of τ
		128	256	512	1024	2048			
Rigid Partitions (Tile, Brick, Concrete, etc.)—*Continued*									
Clinker concrete, 3 in., plastered both sides....	31.	28.	33.	40.	50.	57.	6	40	0.00010
Concrete, 3 in.; 0.75 in. cement and linoleum...	49.0	36.5	37.5	44.5	54.0	65.0	5 §	48	0.000016
Wood Studs and Plaster, Metal Channel Iron and Plaster, etc.									
Wood studs, 2 in. by 4 in., 17 in. o.c., 0.25 in. by 1.5 in. wood lath, 0.375 in. apart, gypsum scratch, lime brown, smooth finish..................	....	......	49.5	....	42.6	52.2	1 *	44	0.000040
Wood stud; wood lath, 0.25 in. by 1.5 in. spaced 0.375 in.; gypsum scratch coat, 0.25 in.; brown coat, 0.25 in.–0.375 in.; finish coat....	18.6	24.4	25.6	29.1	32.2	35.7	3	37	0.00020
Wood studs, etc., as above except lime plaster.....	18.0	27.5	28.8	38.1	46.6	42.9	3	43	0.000050
Wood studs, 2 in. by 4 in., 0.5 in. Celotex, gypsum plaster................	12.0	17.7	24.7	37.0	43.7	36.7	3	40	0.00010
Floor and Ceiling Partitions									
Concrete flat slab floor construction, reinforced. Insulite furred out, applied as ceiling.........	54.4	50.9	54.8	58.7	56.5	53.2	1 *	51	0.0000080
Wood joists. Lower side plastered on wood lath; upper side, subflooring and 0.375 in. finish flooring..................	....	47.9	46.8	40.7	50.1	48.8	1 *	43	0.000050
Wood joists, etc., as above, with floating floor consisting of nailing strips rough and finish flooring	....	57.6	57.5	54.8	62.4	57.6	1 *	53	0.000005
Double Walls									
Tile, double 2 in. solid gypsum, unplastered, unbridged, 2 in. separation; structurally separated..	20.4	25.2	34.2	44.5	51.0	62.6	3	48	0.000016
Tile, etc., as above, except bridged at middle......	20.4	21.3	32.7	37.0	45.6	52.0	3	44	0.000040
Tile, etc., as above, filled with sawdust..........	23.0	21.6	28.1	39.3	47.0	54.0	3	44	0.000040
Tile, double 2 in. solid gypsum, unplastered, unbridged, 4 in. separation; structurally separated..	20.4	28.4	47.4	54.2	59.0	56.8	3	51	0.000008

* Reduction factors from the Bureau of Standards, for the most part, are for the following frequency bands: 150 to 157, 250 to 285, 500 to 547, 1000 to 1070, and 2000 to 2175 cycles. The data for the several frequency bands are recorded under the frequencies to which the frequency bands most nearly correspond.

† For frequencies of 300, 500, and 1000 cycles.

‡ Average value from 128 to 2048 cycles.

§ Berg and Holtsmark's reduction factors are for frequency bands of 100–200, 200–400, 400–800, 800–1600, and 1600–3200 cycles. Their values are recorded under 128, 256, 512, 1024, and 2048 cycles, respectively.

Authority: (1) Bureau of Standards, (2) V. O. Knudsen, (3) P. E. Sabine, (4) Davis and Littler, (5) Berg and Holtsmark, and (6) National Physical Laboratory.

For other methods of measuring transmission coefficients see Knudsen, *Architectural Acoustics*, Chapter XI.

The results of sound-insulation measurements obtained in different laboratories are grouped in Table 2. Each group contains the data for materials or partitions having a number of properties in common. The data in each group have been arranged so as to keep together those from each laboratory, which is necessary since the results on the same panel tested in different laboratories are not always in good agreement. The data from most laboratories give what has been called the *reduction factor* for the panel or partition tested. This *reduction factor* is usually the ratio of the intensities of sound on both sides of the test panel, or, in decibels, is 10 times the logarithm of this ratio. Since the transmission coefficient depends upon the size of the panel and the amount of absorption in the test room, as is shown by eq. (4), the reduction factors published by several laboratories (before about 1935) do not agree with the T.L., which in decibels is $10 \log_{10} 1/\tau$. The compiler has made an attempt to adjust the data from different laboratories in such a manner as to give comparable ratings for all the materials and partitions listed in the tables.

In the table, the data given under "Reduction Factors in Decibels" are the results published by the authors. The data given in the last two columns in the table give the compiler's estimate of the probable average value of the T.L., and the probable average value of the transmission coefficient τ. The *reduction factors* at the different frequencies are useful since they describe the insulation value of the panel or partition at these frequencies. This is often an important matter in the selection of materials or partitions for sound insulation. For example, partitions which have a relatively low insulation value in the frequency range from 500 to 1000 cycles would not be suitable for the insulation of traffic noise and of most noises met in buildings, since most such noises contain a relatively large amount of sound energy in this frequency range. (See Fig. 1.) For this reason, the panels which show the highest values of T.L. may not supply the greatest amount of sound insulation for all types of noise. It is necessary, in determining the best type of partition for each problem which arises in sound insulation, to give consideration to the reduction factors at the different frequencies as well as to the average value of T.L. or τ.

Partitions which possess outstanding merit with regard to both insulation value and practicability are designated by bold-face type in the column which gives the name and description of the partition.*

29. PRACTICAL CONSIDERATIONS IN THE SELECTION OF MATERIALS AND TYPES OF STRUCTURE FOR INSULATION IN BUILDINGS

Since errors as large as 3 or 4 db are inherent in many of the data on sound insulation, small differences in the tabulated results of the preceding article should not be regarded as having much significance. A consideration of the tabulated data suggests the following generalizations concerning the insulative properties of different building materials and partitions:

1. The insulation value of rigid masonry or monolithic partitions increases directly as the logarithm of the weight per square foot of wall section—so that the rate of increase of insulation with increased weight is relatively slow for partitions which are heavier than, say, 30 or 40 lb per sq ft. As a consequence, it is often in the interest of both insulation and economy to substitute two or more light-weight partitions, or specially composed partitions, for heavy masonry partitions. There are many occasions in building practice where this may be done with a gain in insulation, with a reduction in the dead load of the building, and at a reduced cost. However, for thin partitions, where the dead load is not a serious problem, dense rigid panels such as plastered brick or solid tile provide a satisfactory means of obtaining a T.L. of about 40 to 45 db.

2. Lime plaster on wood lath and wood studs gives better insulation than an equal thickness of gypsum plaster on wood lath and wood studs. Sabine reports an advantage of about 9 db for lime plaster as compared with gypsum plaster, and the Bureau of Standards reports an even greater difference. Wood stud and plaster partitions rate slightly higher than channel iron and plaster.

3. Wood partitions, with tongue-and-groove joints, provide more insulation than masonry partitions of the same weight or thickness. (The development of cracks in wood partitions, however, will greatly reduce their insulation value.)

* More extensive tables on sound-insulation data will be found in Knudsen's *Architectural Acoustics*, Glover's *Practical Acoustics for the Constructor*, Constable and Ashton, Sound Insulation of Single and Complex Partitions, *Philosophical Mag.*, Vol. 23, 161–181 (1937), and Knudsen and Harris, *Acoustical Designing in Architecture* (1949).

4. Double partitions seem to offer the most feasible means of obtaining high insulation at a reasonable cost and reasonable dead load. The separate partitions should be as completely insulated from each other as possible, as the introduction of structural ties between the separate partitions tends to convert the two partitions into a single rigid one and thus greatly reduces the insulation. The suspension of a blanket or fiber board between double partitions or between the wood studs or channel irons in staggered stud partitions is often

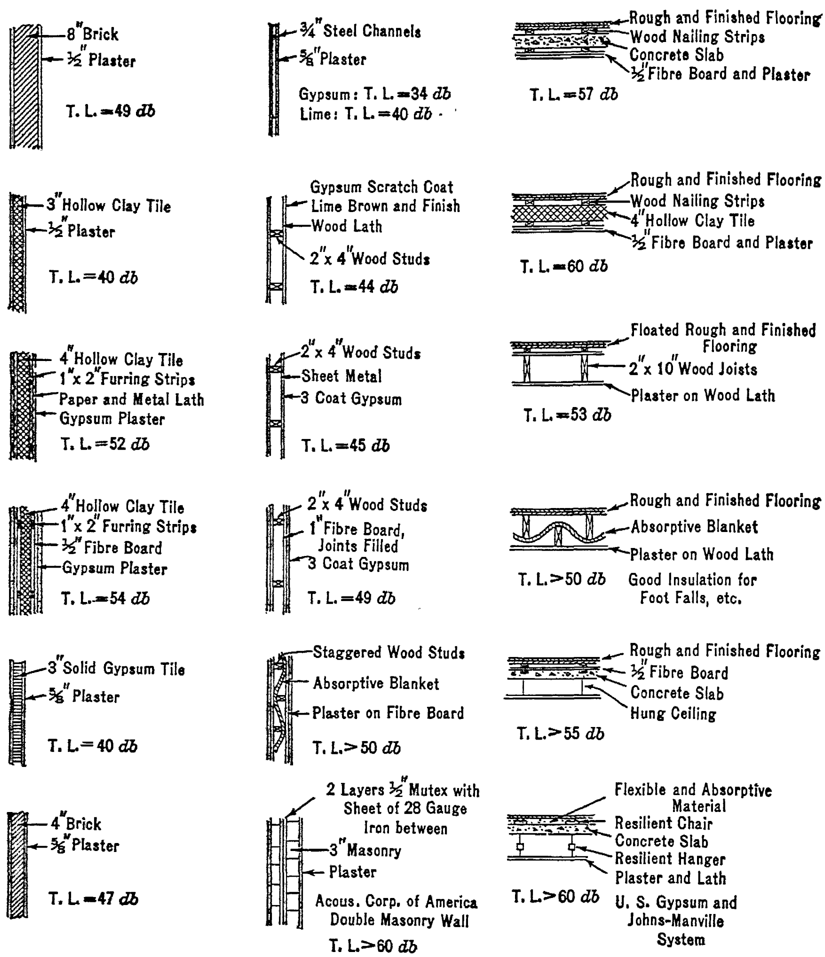


FIG. 8. Recommended Types of Structure for Sound Insulation

an aid to insulation. The addition of any absorptive material in the air space between double partitions contributes considerably to sound insulation unless the absorptive material makes a rigid or semirigid *bridge* between the two partitions, in which case it may be worse than nothing. Thus, the addition of cinders, pumice, or other rigid-porous materials between structurally separated partitions will sometimes reduce the overall insulation.

A number of satisfactory types of construction for obtaining wall partitions having a T.L. greater than 40 db and floor and ceiling partitions having a T.L. greater than 50 db are indicated in Fig. 8. These methods of construction will meet most of the requirements for sound insulation that will arise in connection with the design of buildings. The average T.L. is given for each partition. (See also references given on pp. 12-67 and 12-75.)

30. CALCULATION OF INSULATION IN BUILDING DESIGN

From a simple consideration of the transmission of sound through the boundaries of a room it can be shown that the *noise-reduction factor* for a room (the number of decibels the intensity of the sound is reduced by transmission through the boundaries) is given by

$$\text{Noise-reduction factor} = 10 \log_{10} \frac{a}{T} \tag{5}$$

where a is the total absorption of the room, and $T = \tau_1 A_1 + \tau_2 A_2 + \tau_3 A_3 + \ldots$ is the total transmittance for the boundaries of the room. τ_1, τ_2, and τ_3 are the coefficients of transmission of the different parts of the boundaries of the room, and A_1, A_2, and A_3 are the corresponding areas of these boundaries.

In order to illustrate the use of eq. (5) a typical calculation will be made for determining the noise-reduction factor of a small studio.

Volume of room = 50,000 cu ft.
Total absorption in room, including audience = 2400 sabins.

Description of Walls, Ceiling, Windows, and Doors; and the Transmittance through These Surfaces

Material	Area, A, sq ft	τ	τA
4-in. concrete slab ceiling plus 1/2 in. acoustical material	2500	0.000025	0.0625
8-in. brick walls plus 1/2 in. acoustical material	4500	.0000080	.0360
3/16-in. glass windows, closed	400	.00110	.440
1 1/2-in. hardwood doors, good closure	100	.00031	.031
Total transmittance (T)	....	....	0.5695

Therefore

$$\frac{a}{T} = \frac{I_1}{I_2} = \frac{2400}{0.5695} = 4210$$

and

$$10 \log_{10} 4210 = 36.2 \text{ db}$$

That is, the noise-reduction factor, or the effective insulation which the room provides against outside noise, is 36.2 db. Thus, if the studio is located where the outside noise is at a level of 60 db, the level of the noise which reaches the studio will be 60 − 36.2 or 23.8 db. It will be noted that most of the transmitted noise is that which comes through the glass windows, and that therefore the noise-reduction factor can be increased considerably by means of double windows or by dispensing with the windows.

ACOUSTIC DESIGN OF AUDITORIUMS

By Vern O. Knudsen

31. THE HEARING OF SPEECH IN AUDITORIUMS

Four principal factors affect the hearing of speech in auditoriums: the shape of the room, the loudness of the speech which reaches the listeners, the reverberation characteristics of the room, and the amount of noise in the room. If average speech is loud and distinct, and entirely free from the interfering effects of noise and reverberation, the percentage articulation for the average listener will be 96, that is, 96 out of 100 meaningless monosyllabic speech sounds will be heard correctly.

In an equation for calculating the percentage articulation for a room it will be necessary to introduce reduction or distortion factors due to (1) the shape of the room, (2) inadequate loudness, (3) excessive reverberation, and (4) extraneous noise. It is possible to represent approximately the percentage articulation in any room by the equation

$$\text{Percentage articulation} = 96 k_s k_l k_r k_n \tag{1}$$

where k_s = the reduction factor due to the shape of the room.
k_l = the reduction factor due to the inadequate loudness of speech.
k_r = the reduction factor due to the excess of reverberation in the room.
k_n = the reduction factor due to the extraneous noise in the room.

In the ideal room each of these factors will be equal to unity, so that the percentage articulation under such conditions would be 96 per cent.

Experimental data have been obtained by means of which it is possible to determine the appropriate values of these four factors for any auditorium, although considerable research remains to be done before accurate values of k_s are known. When these factors are determined for a certain auditorium and substituted in eq. (1) the resulting product gives the probable average percentage articulation in that auditorium.

It is admittedly only an approximation to represent these factors by single numbers. For example, the reverberation is described not by a single number but by a curve giving the times of reverberation at different frequencies. However, if one takes the time of reverberation for a tone of 512 cycles, one will have a fairly reliable index for representing the condition of reverberation in a room, especially if the reverberation time does not vary too widely at different frequencies. Such a single index is very useful for rating the acoustical quality of rooms. In problems of design, on the other hand, consideration should be given to the reverberation throughout the entire range of frequencies. Similarly, the noise spectrum in the room and the shape of the room cannot be represented rigorously by single numbers. However, in the rectangular room of conventional shape it is not only admissible to use a single factor to represent the effect of shape on articulation, but the factor k_s will not deviate appreciably from unity.

The four factors which affect the hearing of speech in rooms will now be considered separately.

SHAPE. It is not only necessary to avoid shapes which will produce acoustical defects, such as echoes, flutters, sound foci, interfering effects, and "whispering gallery" effects, but it is of prime importance to design shapes which will facilitate the most advantageous flow of sound energy to all auditors in the room.

There are three outstanding forms which should never be tolerated: (1) those which will produce a pronounced focusing of sound, thus giving an excessive concentration of sound in some places and a scarcity of sound in other places; (2) those which will produce excessive delays between the sound which reaches the auditors by a direct path from the source and that which reaches the auditors by reflection from the ceiling or walls; and (3) those in which the sound reaching the auditors travels a relatively long distance, at grazing incidence, over a highly absorptive surface. The sound which comes by the reflected paths always has to travel a greater distance than that which comes by the direct path, and if the difference in these path lengths is as great as 65 ft the reflected sound will be delayed to the extent that it is heard as a separate sound; that is, the delayed sound produces an echo. Even when the reflected sound is delayed as much as 50 ft it unites with the direct sound sufficiently out of phase to produce a masking or *blurring* interference. Figure 1 exhibits a characteristic defect which results from the use of concave surfaces.

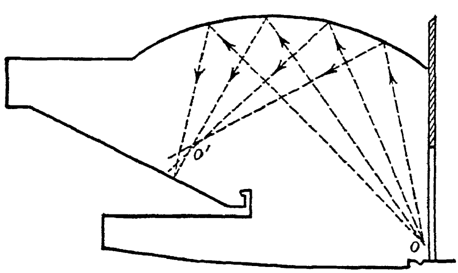
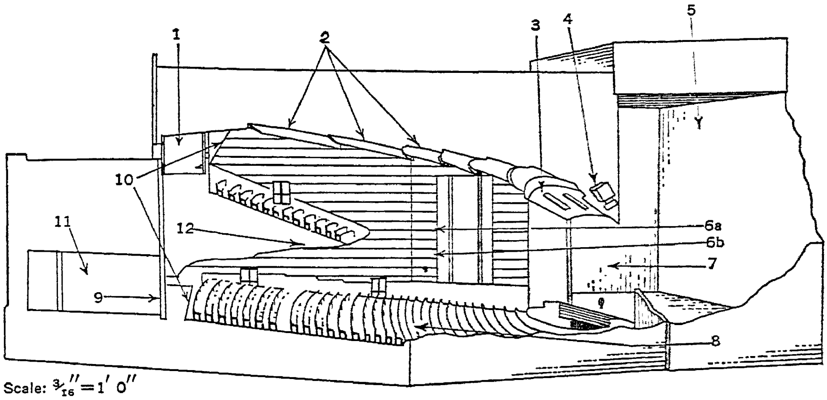


Fig. 1. Reflection of Sound from a Domed Ceiling

Figure 2 shows a longitudinal section of a cut-away model of an auditorium which is not only free from concentrations, dead spots, and interfering reflections but also is so shaped as to give a nearly uniform distribution of diffusely reflected sound to all parts of the auditorium, with a slight preference for the more remote parts.

In many auditoriums, and even in some sound-recording studios, it may be difficult or even impossible to avoid large and therefore troublesome differences of path between the direct and reflected sound. In such instances it is advisable to break up the surfaces producing these delayed reflections by introducing coffers, beams, pilasters, or other irregularities in contour. A number of rooms, highly acclaimed for their good acoustics, have been designed with walls and ceiling deliberately covered with polycylindrical sound diffusers. Figure 3 is a typical example. In rooms where public-address systems are required, the architect will have greater freedom in designing the shape and size of the room.

Both speech and music rooms should be designed so that the auditors receive a relatively large amount of sound which travels directly from the source or from reflectors located sufficiently near the source so that this reflected sound is nearly in phase with the direct sound. The stage, pulpit, platform, choir loft, or other location for the speaker or performer should be well elevated above the audience and provided with large, reflective



Scale: $\frac{3}{16}'' = 1'\ 0''$

1. Acoustically treated projection booth with sound amplification equipment and controls for sound monitoring.
2. Ceiling planes reflect sound to all parts of the auditorium.
3. Three-channel public address system to reproduce stage sound in "auditory perspective".
4. High-fidelity speakers: bass-compensated dynamic speaker for low tones; high frequency directional horns for high tones.
5. Backstage treated with acoustical plaster to reduce "stage echoes".
6. Acoustic treatment on walls: over-all distribution in alternate bands of (a) acoustic tile and (b) hardwall plaster.
7. Proscenium splays. horn-like shape of stage opening projects sound to audience.
8. Upholstered seats: absorption value of each seat equivalent to that of a person's clothing.
9. Double doors to foyer insulate against external noises.
10. Slanting rear walls on main floor and balcony reflect sound down toward rear seats.
11. Acoustically treated foyer to reduce external noises.
12. Streamlined balcony improves flow of sound to rear seats.

FIG. 2. Cut-away Model of the High School Auditorium for Whittier, California. (William Harrison, Architect.)

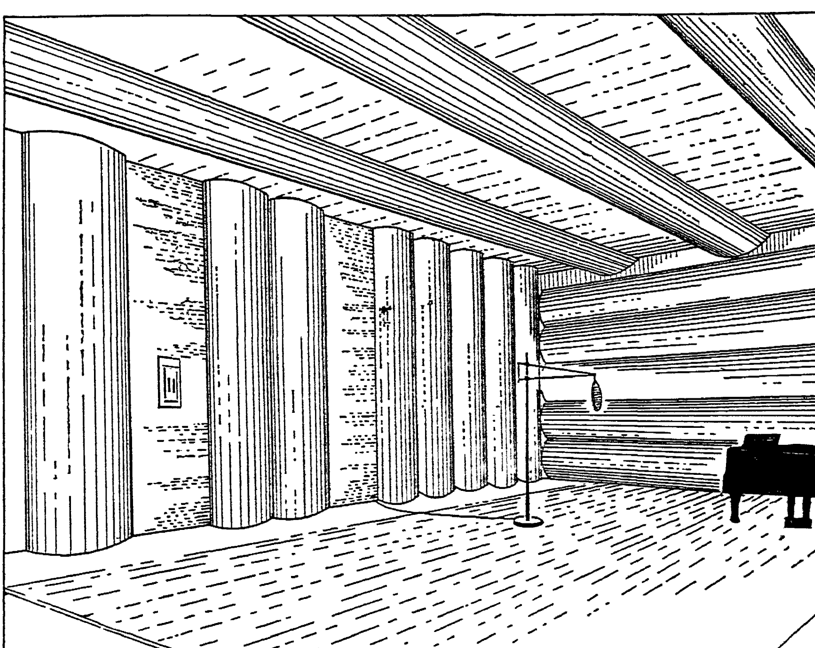


FIG. 3. Polycylindrical Sound Diffusers in an RCA Disk Recording Studio in New York. (Volkmann.)

surfaces located behind, above, and on both sides of the position where the sound originates. In addition, especially in large rooms, the floor should rise progressively toward the rear of the room, or the room should be designed with one or more balconies, so that all auditors obtain an abundance of direct or beneficially reflected sound—the path length of the reflected sound should not exceed that of the direct sound by more than 50 ft, and preferably less than that, so that the reflected sound will reinforce, and not interfere with, the direct sound. Long rooms with level floors, with a low platform or no platform at all, and with a highly absorptive ceiling, are especially poor speech rooms, since much of the sound reaching the auditors in such rooms travels at near grazing incidence over the highly absorptive audience and also along the absorptive ceiling. (See Fig. 4.) The intensity of free progressive sound waves propagated over such surfaces has been found to diminish as the inverse fourth power of the distance from the sound source. Properly designed reflective surfaces, large compared with the wavelength of the sound, will largely compensate for such excessive losses and will insure adequate loudness of unamplified speech for rooms having volumes of less than about 50,000 cu ft. For larger rooms it is necessary to amplify the speech with a suitable public-address system.

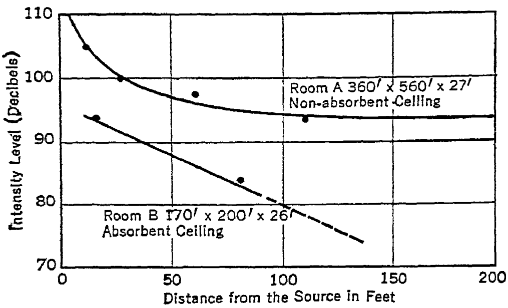


FIG. 4. Variation of Intensity of Sound with Distance from Source in Large Rooms with Absorbent and Non-absorbent Ceilings, Showing the Excessive Decrease in Level Which Occurs When Sound Travels over or along an Absorptive Surface

For further details regarding shape of rooms consult Bagenal and Wood, *Planning for Good Acoustics;* Knudsen and Harris, *Acoustical Designing in Architecture;* and the current and bound volumes of the *Journal of the Acoustical Society of America.*

NOISE. The curve in Fig. 5, obtained empirically, gives the value of the noise-reduction factor k_n for different amounts of noise. The abscissa is the ratio of the sound level of the noise, in decibels, to that of the speech, also in decibels. Thus, when the noise

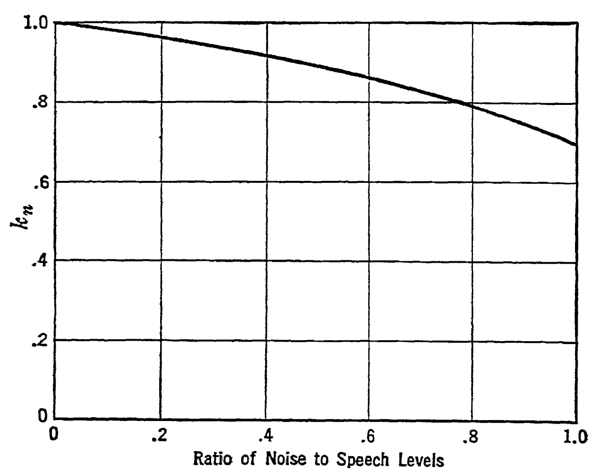


FIG. 5. Effect of Noise on the Hearing of Speech

is at the same level as the speech, the abscissa is 1.0. Although the value of k_n given in these curves is only an approximation, based upon a limited number of measurements, experience has indicated that it is useful in problems of design or correction.

LOUDNESS. The curve of Fig. 3, p. 12-31, gives the percentage articulation for speech as a function of the sound level of speech, as determined by Fletcher and Steinberg. Thus, the optimal sound level of speech, in quiet surroundings, is 70 db above threshold, and for levels below 40 db the articulation drops off rapidly. The loudness reduction factor k_l to be used in eq. (1) is obtained by dividing the percentage articulation by 96. The curve in Fig. 6 gives the average speech power, in microwatts, of the average speaker in rooms of different size. By means of this curve, and the amount of absorption in the

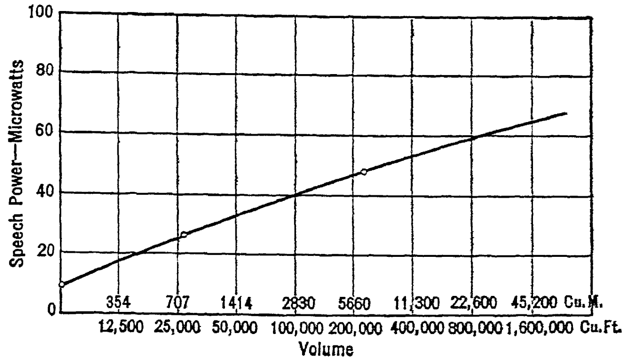


FIG. 6. Average Speech Power of Speakers in Rooms of Different Sizes

room, it is possible to calculate the average sound level of speech for any auditorium, and then from Fig. 3, p. 12-31, to calculate the appropriate value of k_l. In Table 1 are given values of k_l for rooms of different size and different times of reverberation.

Table 1. Values of k_l for Use in Eq. (1)

Time of Reverberation, sec	Volume of Room, cu ft							
	12,500	25,000	50,000	100,000	200,000	400,000	800,000	1,600,000
0.50	0.96	0.94	0.92	0.90	0.88	0.85	0.81	0.76
0.75	.97	.95	.93	.91	.89	.87	.84	.80
1.00	.97	.96	.94	.92	.90	.88	.86	.82
1.25	.97	.96	.95	.93	.91	.89	.87	.83
1.50	.98	.96	.95	.94	.92	.90	.88	.84
2.00	.98	.97	.96	.95	.93	.91	.89	.86
3.00	.98	.97	.97	.96	.94	.92	.91	.88
4.00	.98	.98	.97	.96	.95	.94	.92	.89
6.00	.99	.98	.98	.97	.96	.95	.93	.91
8.00	.99	.99	.98	.97	.96	.95	.94	.92

REVERBERATION. The curve shown in Fig. 7, empirically determined, gives the value of k_r, the reverberation reduction factor, for different times of reverberation at 512 cycles. This curve is based upon the results of speech articulation data obtained in four-

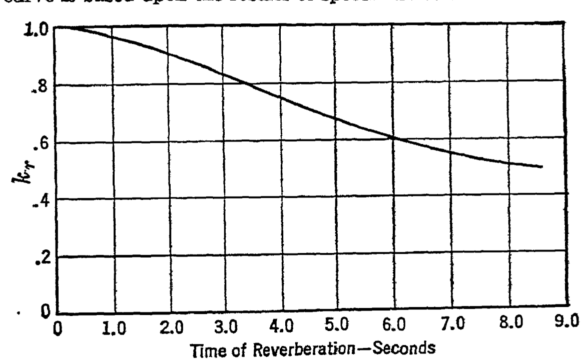


FIG. 7. Effect of Reverberation on Hearing of Speech

teen different auditoriums having times of reverberation varying from less than 1 sec up to more than 8 sec. It is seen that k_r decreases almost uniformly as the time of reverberation increases from 1 to about 6 sec. The abscissa in this curve gives the reverberation time at 512 cycles. In general, the time of reverberation at 128 cycles should be about 25 to 50 per cent longer than it is at 512 cycles, and the reverberation time should remain approximately constant at frequencies above 512 cycles, increasing slightly for frequencies above 2048 cycles.

COMBINED EFFECTS OF LOUDNESS AND REVERBERATION. Since the addition of absorption to a room diminishes the loudness of speech as produced by the average speaker, it is reasonable to assume that there will be an optimal time of reverberation for speech rooms. This optimal time will be attained when a further reduction in the reverberation will concurrently reduce the loudness of speech to the extent that the impairment produced by the diminished loudness will just compensate for the improvement occasioned by the reduction of the reverberation. The curves shown in Fig. 8 were calculated by means of eq. (1), using the appropriate values of the loudness-reduction and reverberation-reduction factors. These curves give approximately the average percentage articulation

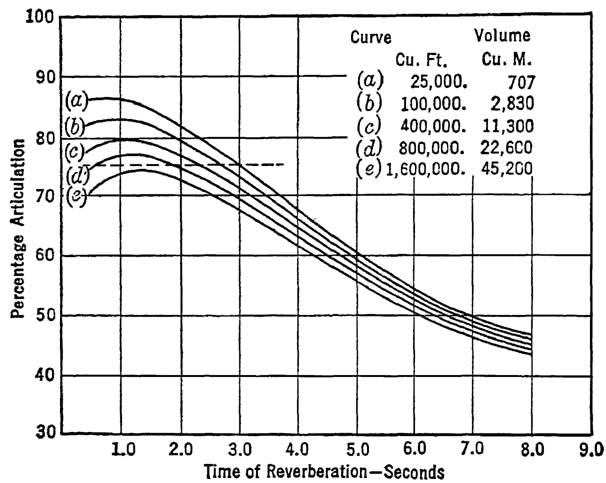


FIG. 8. Percentage Articulation Curves for Rooms of Different Sizes and Different Times of Reverberation

which will obtain in rooms of different sizes and different times of reverberation, for the average speaker, without artificial amplification. If an articulation of 75 per cent is regarded as the minimal for satisfactory hearing (see also article 29), it is apparent that the average speaker will not be heard satisfactorily in an auditorium larger than about 1,000,000 cu ft, no matter what condition of reverberation has been provided for the room. Also, in an auditorium having a volume of 100,000 cu ft the time of reverberation should not exceed 2.70 sec. The curves in Fig. 8 are based upon the assumption that the noise level has been reduced to 30 db, an unusually quiet room, and that $k_s = 1.0$. The need for artificial amplification of speech in large rooms is apparent; amplification should be provided in all rooms larger than about 50,000 cu ft, and even in smaller rooms when considerable noise is present.

The curves in Fig. 8 apply to the average speaker. Speakers with weak voices will not be heard so well, as is indicated by the curves, and speakers with loud voices will be heard better.

In all the curves in Fig. 8 it will be seen that the articulation is a maximum for a particular time of reverberation, and that this optimal time of reverberation increases with the size of the room, a fact which is well established by experience. Furthermore, the percentage articulation in Fig. 8 is the average value; in general, the articulation diminishes progressively with the increase of distance from the speaker.

32. MUSIC ROOMS

The reverberatory properties of a room are of even greater significance for music than they are for speech. The acoustical properties of a music room are no less important than those of the musical instrument to be played in that room; indeed, the room and instrument together comprise a coupled system, and it is this combined system that the ear or microphone "hears." The resonant frequencies of a room, considered in article 20, depend on the dimensions of the room; their intensities and their rates of growth and decay are largely influenced by the distribution of the absorptive and reflective materials over the boundaries of the room.

A music room should be so dimensioned, shaped, and treated with absorptive and reflective materials as to support and enhance the rich quality of the individual tones and harmonies of music, and to join together these separate tones and harmonies so that they coalesce into a continuously flowing melody. The best music rooms, like the best violins, are those which are free from prominent resonances and which have a relatively uniform steady-state response throughout the entire frequency range. Rooms in which the ratio of height, width, and length is approximately 2 : 4 : 5, and in which furred-out wood paneling and wood flooring on wood joists comprise most of the interior boundaries, are found to meet these conditions and usually are highly acclaimed by both performers and listeners. The use of non-parallel pairs of opposite walls, of polycylindrical diffusers of wood veneer, similar to those illustrated in Fig. 3, and of "patches" of absorptive materials distributed over walls and ceiling so as to give an ergodic diffusion of sound has given good results.

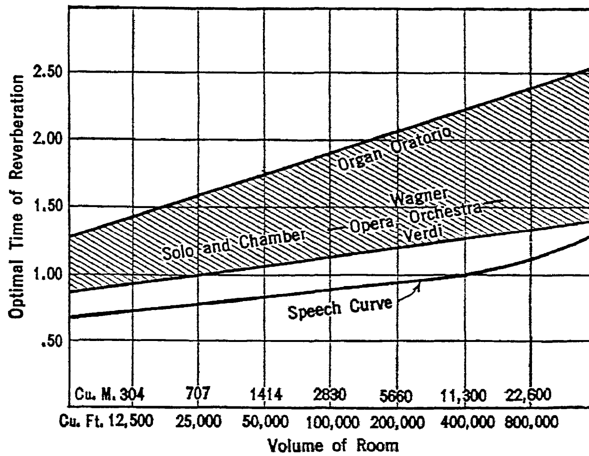


FIG. 9. Optimal Reverberation Times for Speech and Music Rooms

The design of music rooms should always be guided by the principles of wave acoustics. However, the optimal reverberation characteristics can be best calculated by means of the approximate formulas of geometric acoustics, such as eq. (3) of article 19.

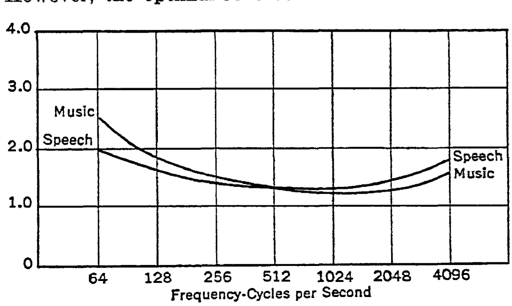


FIG. 10. Optimal Reverberation Times at Different Frequencies, for Speech and for Music, When the Optimal Reverberation Time at 512 Cycles Is 1.3 Seconds. Similar curves should be used when the optimal time at 512 cycles differs from 1.3 seconds.

The optimal reverberation time for music rooms depends not only on the size of the room but also on the type of music to be performed in the room. The ideal arrangement should provide for adjustable reverberation so that the optimal reverberatory properties can be readily obtained for all musical performances for which the room is designed.

The chart in Fig. 9 shows the optimal times of reverberation for both speech rooms and music rooms. This chart applies for a frequency of 512 cycles. As for speech, the reverberation time at 128 cycles should be approximately 25 to 50 per cent longer than the time for 512 cycles; the reverberation time should remain approximately constant for frequencies between 512 and 2048 cycles, and should increase slightly for frequencies above 2048 cycles (see Fig. 10).

Experience has shown that the most satisfactory reverberation time for radio broadcasting or sound-recording studios is about two-thirds to three-fourths of the accepted time for speech or music rooms (see also Section 16, Sound-reproduction Systems).

33. PRACTICAL PROCEDURE FOR OBTAINING GOOD ACOUSTICS IN BUILDINGS

The procedure for obtaining good acoustics in buildings begins with the selection of the site and ends with the furnishing, testing, and maintaining of the building. The necessary steps, approximately in chronological order, are as follows:

1. *Selection of a suitable site* (sound studios, theaters, schools, churches, and hospitals, especially, should be located in quiet surroundings).
2. *Noise survey*, at the proposed site, to determine the amount of insulation required to reduce the noise level in the building to a satisfactory point.
3. *Insulation against outside noise*, which includes not only the selection of the proper sound-insulative and sound-absorptive constructions but also the proper arrangement of rooms, corridors, entrances, windows, landscaping, and other appurtenances of the building.
4. *Design of the shape of the room* (shapes should be designed which not only will avoid such acoustical defects as echoes, interfering reflections, room flutter, and sound foci, but also will facilitate the most advantageous flow of diffuse sound energy to all auditors in the room, and at the same time will preserve or even enhance the natural beauty of speech and music).
5. *Control of the noise within the building*, including solid-borne as well as air-borne noise and vibration.
6. *Selection and distribution of the absorptive and reflective materials to provide the optimal conditions for both steady-state and transient sounds throughout the room*, a problem which deserves special study and careful planning, and one which involves, besides the acoustical characteristics of the materials, such properties as structural strength, decorative possibilities, adaptability to the surfaces available for, or requiring, absorptive treatment, maintenance, sanitation, ease of application, fire hazard, absorption of water, attraction for vermin, "fool-proofness," durability, and cost.
7. *Supervision of the installation of acoustical materials* (especially necessary for the application of acoustical plaster—in large buildings it is advisable to require the plastering contractor to prepare a small room for test and approval before the plaster is used in other parts of the building).
8. *Installation of high-quality amplifying equipment under the supervision of a competent engineer* is necessary in all large auditoriums; even in rooms seating as few as 200 or 300 persons it will be found that many speakers have weak voices that require amplification.
9. *Inspection of the finished building* should include tests to determine whether the sound insulation, the sound absorption, and the other acoustical properties have been satisfactorily attained.
10. *Maintenance instructions, preferably in writing, should be left with the building manager*, indicating (*a*) how the acoustical materials can or cannot be cleaned or redecorated, (*b*) which furnishings in the building are essential to good acoustics, and (*c*) how the humidity of large speech and music rooms should be maintained in order to avoid excessive absorption of high-pitched sounds.

The foregoing steps, or their equivalent, if carefully executed, will lead to good acoustics. Developments in modern theories of room acoustics, supplemented by additional empirical data, will contribute to more reliable criteria than are now available for determining the best acoustical shape of a room and the most favorable distribution of absorptive and reflective materials throughout the room; but if proper use is made of what is now known there need be no anxiety respecting the outcome in the acoustics of buildings—the outcome will be good.

BIBLIOGRAPHY

Bagenal and Wood, *Planning for Good Acoustics*. Methuen, London (1931).
Davis and Kaye, *The Acoustics of Buildings*. G. Bell and Sons, London (1927).
Fleming and Allen, *Modern Theory and Practice in Building Design*, The Institution, London (1945).
Fitzmaurice and Allen, *Sound Transmission in Buildings*, His Majesty's Stationery Office, London (1939).
Glover, *Practical Acoustics for the Constructor*. Chapman and Hall, London (1933). (Contains a complete bibliography of books and journal articles.)
Knudsen, V. O. *Architectural Acoustics*. John Wiley (1932).
Knudsen and Harris, *Acoustical Designing in Architecture*. John Wiley (1949).
Rettinger, *Applied Architectural Acoustics*, Chemical Publishing Co. (1947).
Sabine, P. E., *Acoustics and Architecture*. McGraw-Hill (1932).
Strutt, Raumakustik. *Handbuch der Experimentalphysik*, Bank XVII/2. Akad. Verlagsgesellschaft, Leipzig (1933).
Watson, F. R., *Acoustics of Buildings*. John Wiley (1938).

SECTION 13

ELECTROMECHANICAL-ACOUSTIC DEVICES

ELECTROMECHANICAL-ACOUSTIC DEVICES

EFFECTS OF THE ACOUSTIC MEDIUM

By Hugh S. Knowles

"An electroacoustic transducer is a transducer which is actuated by power from an electrical system and supplies power to an acoustic system, or vice versa." (I.R.E. Standards.)

The theory of operation of electroacoustic transducers is an extension of the theory of electromechanical transducers, in which account is taken of the reaction of the fluid, or acoustic medium, on the diaphragm. (See Section 5, article 33.)

1. PHYSICAL PROPERTIES OF COMMON ACOUSTIC MEDIA

The velocity of a sound of small (infinitesimal) wave amplitude depends on the elasticity and density of the fluid. The velocity of propagation, c, of a sound wave is

$$c = \sqrt{\frac{k}{\rho}} = \sqrt{\frac{\gamma p_0}{\rho}} \quad \text{cm sec}^{-1} \tag{1}$$

where k is the volume modulus of elasticity, ρ is the density in grams cm^{-3}, γ is the ratio of the specific heat at constant pressure to that at constant volume, and p_0 is the static pressure of the fluid.

The characteristic impedance, z_0 (also sometimes called surge impedance or acoustic resistance), of the medium is

$$z_0 = \sqrt{k\rho} = \rho c \quad \text{mechanical ohms cm}^{-2} \tag{2}$$

1. Air. The density, ρ, at 20 deg cent and $p_0 = 760$ mm ($\doteq 10^6$ dynes cm^{-2}) is 0.001205 gram cm^{-3}; $\gamma \doteq 1.41$, giving

$$c \doteq 33{,}060 + 61\theta \quad \text{cm sec}^{-1} \tag{3}$$

where θ is the temperature in degrees centigrade. Also the radiation resistance of air is

$$z_0 = 42.8 - 0.079\theta = 41.2 \quad \text{mechanical ohms cm}^{-2} \text{ at 20 deg C} \tag{4}$$

The pressure level, L_p, in decibels is

$$L_p = 20 \log_{10} \frac{p}{p_0} = 74 + 20 \log_{10} p \tag{5}$$

where p is the pressure and p_0 the reference pressure of 0.0002 dyne cm^{-2}. The intensity, I, in the direction of propagation, of a plane or spherical "free" (i.e., no reflections) sound wave is

$$I = \frac{p^2}{\rho c} = 2.42 \times 10^{-9} p^2 \quad \text{watt cm}^{-2} \tag{6}$$

where p is the rms sound pressure in dynes cm^{-2}.

2. Hydrogen. The density, ρ, of hydrogen at 0 deg cent and at $p_0 = 760$ mm is approximately 0.00009 gram cm^{-3}. The velocity $c \doteq 1.26 \times 10^5$ cm sec^{-1}, from which $z_0 = 11$ mechanical ohms cm^{-2}.

3. Water. The density, ρ, of water is $= 1.0$ gram cm^{-3}. The value of c at 20 deg cent $= 1.46 \times 10^5$ cm sec^{-1}, from which $z_0 = 1.46 \times 10^5$ mechanical ohms cm^{-2}.

REACTION OF ACOUSTIC MEDIUM ON A DIAPHRAGM. The audible frequency range of sounds covers roughly 10 octaves. Even the more important range from 80 to 8000 cycles covers nearly 7 octaves, giving a wavelength range of roughly 428 cm (14.1 ft) to 4.28 cm (1.69 in.). Radiation, diffraction, and reflection phenomena which depend on the relative length of the sound wave and the linear dimensions of the radiator, or collector, therefore differ greatly in different portions of the wavelength range, and hence simplifying assumptions can ordinarily be made only over restricted frequency ranges.

2. MECHANICAL IMPEDANCE TO MOTION OF SOME SIMPLE ACOUSTIC RADIATORS

The impedance to motion of a diaphragm is altered by its contact with the acoustic medium or fluid. If the diaphragm has an effective area S and a mechanical impedance z_m, its total effective mechanical impedance is $z_{mf} = z_m + z_f$, where z_f is the fluid impedance (see also Section 5, article 33).

SOURCE RADIATING PLANE SOUND WAVES. The impedance per unit area in contact with the fluid is the characteristic impedance of the medium, and the total increase in mechanical impedance to motion of a radiator of effective area, S, resulting from contact with the fluid medium, is

$$z_f = r_f = z_0 S \tag{7}$$

or $\doteq 41.2 \times S$ mechanical ohms in the case of air. At any frequency, the radiated acoustic power is

$$P_f = \dot{s}^2 r_f 10^{-7} \tag{8}$$

or $4.12 \times 10^{-6}\dot{s}^2 \times$ watts in the case of air ($\dot{s}$ is the velocity of the radiator element).

PULSATING SPHERE RADIATING SPHERICAL SOUND WAVES INTO AN UNLIMITED ACOUSTIC MEDIUM. A pulsating sphere is one in which the surface vibrates or pulsates with small amplitude and uniform velocity in a radial direction. At any frequency the impedance per unit area in mechanical ohms is

$$z_A' = r_A' + jx_A' = \rho c\left(1 + \frac{1}{jkR}\right)^{-1} = \rho c\,\frac{k^2R^2}{1 + k^2R^2} + j\omega\,\frac{\rho R}{1 + k^2R^2} \tag{9}$$

where $k = 2\pi/\lambda = \omega/c = 2\pi f/c$, where λ is the length of the emitted sound wave, and R is the radius of the sphere in centimeters. This leads to an equivalent circuit of parallel mass and resistance elements of values (for the whole surface) $m_f = 4\pi\rho R^3$ grams and $r_f = 4\pi R^2 \rho c$ mechanical ohms.

Graphical plots of r_A' and x_A' are given in Fig. 1. The imaginary or reactance term is in phase with the acceleration and is an inertia or mass reactance term.

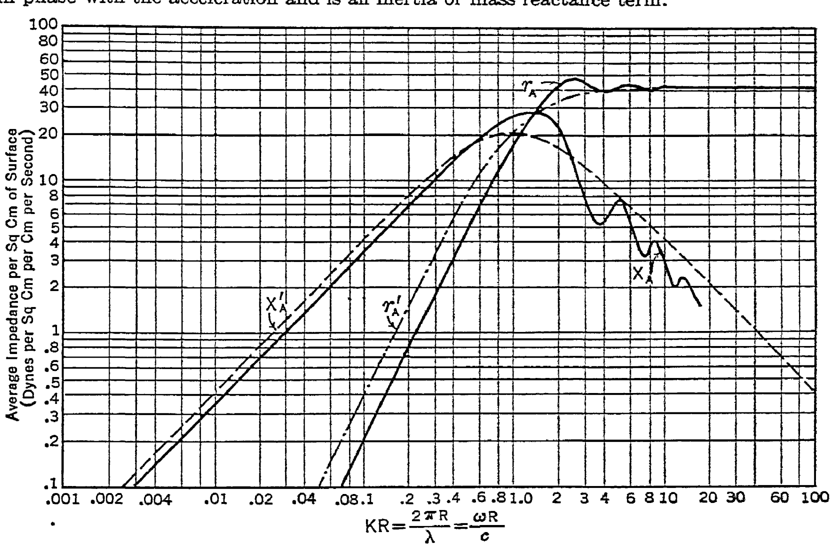


FIG. 1. Air Resistance and Reactance per Unit Area on One Side of a Pulsating Sphere of Radius R (r'_A and x'_A) and on One Side of a Circular Piston of Radius R (r_A and x_A) in an Infinite Baffle

The fluid increases the impedance to motion of the spherical surface or diaphragm by an amount $z_f = Sz_A'$. At any frequency the radiated acoustic power is $P_f = \dot{s}^2 r_f \times 10^{-7}$ watt.

Special Case of a "Point" Source ($R \doteq 0$). When the length of the sound wave is large in comparison with the radius of the sphere ($\lambda \gg R$ and $k^2R^2 \ll 1$),

$$z_f \doteq 4\pi\rho R^4 \frac{\omega^2}{c} + j\omega 4\pi\rho R^3 \quad \text{mechanical ohms} \tag{10}$$

CIRCULAR RIGID DIAPHRAGM OR "PISTON" reciprocating or vibrating sinusoidally in a (perfect-fitting) cylindrical hole, in an infinite plane wall or baffle, and radiating into an unlimited fluid ("semi-infinite" medium) on each side of the baffle, is another configuration of interest. (R is the radius in centimeters.)

The impedance per unit area is not constant over the surface in this case, but since the radiator is assumed rigid an average value may be taken. At any frequency the impedance in mechanical ohms cm^{-2} is given by $z_A = r_A + jx_A$ as plotted in Fig. 1.

Since the imaginary term is in phase with the acceleration, it is a mass reactance term. The increase in mechanical impedance to motion of a piston, $z_f = r_f + jx_f = S(r_A + jx_A)$, which results from an air "load" may be obtained directly from these curves by multiplying the ordinate for any value of kR by the area of the piston. When the fluid is air, piston radiators are usually operated with fluid on both sides of the baffle. Both sides of the piston must then be considered, in which case $S = 2\pi R^2$. The total apparent increase in mass of the piston is $m_f = Sx_A/\omega$. At any frequency, the radiated acoustic power is $P_f = \dot{s}^2 r_f \times 10^{-7}$ watt.

Special Case of a "Point" Source ($R \doteq 0$). When the length of the sound wave is large in comparison with the radius of the piston, the impedance per unit area is

$$z_A \doteq \frac{\rho R^2\omega^2}{2c} + j\omega\frac{8}{3\pi}\rho R \tag{11}$$

Reference to Fig. 1 will show that, for small values of kR, $x_A \gg r_A$, so that $z_A \doteq jx_A$. That is, the increase in impedance to motion of the piston, which results from contact with the fluid, is largely reactive. At this low-frequency condition the radiation impedance of the piston may be represented by an equivalent circuit of parallel mass and resistance elements of values $8\rho R^3/3$ grams and $(128/9\pi^2)\pi R^2\rho c$ mechanical ohms respectively. The values are $0.00656d^3$ grams and $301d^2$ mechanical ohms, d being the piston diameter in inches.

It has been found experimentally that when a conventional rigid conical diaphragm loudspeaker is used S is approximately the area of the "base," or large end, of the cone; also, that the usual magnetic structure changes the radiation from the rear surface of the cone by a negligible amount at low frequencies.

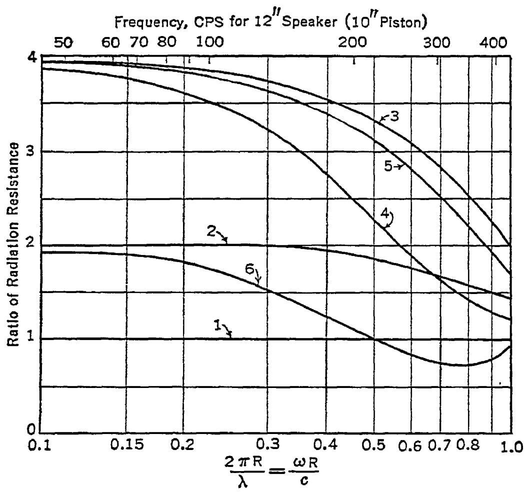


FIG. 2. Ratio of Total Radiation Resistance of a Piston, Vibrating in the Presence of Others, to the Radiation Resistance of a Single Piston. Curve 1, single piston. Curve 2, each of two tangent pistons. Curve 3, each of four tangent pistons in square array. Curve 4, each outer, and Curve 5, each inner, piston of four tangent pistons in a straight line. Curve 6, each of two pistons with centers three diameters apart. All pistons have equal velocities (both magnitude and phase) and equal diameters, and vibrate in an infinite plane baffle.

MULTIPLE PISTONS. When more than one piston radiates into a common region in the medium any one position experiences not only the force on its surface arising from its own vibration but additional forces due to the vibration of the other pistons. The resultant force is the vector sum of the individual forces. The magnitude and phase of each force depend on the diameter, velocity, and frequency of vibration of the piston giving rise to the force and upon its distance from the reference piston. The ratio of the resultant force on any piston, to its velocity, is the total fluid or radiation impedance and comprises the self-impedance due to its own vibration, discussed above, and the mutual impedances due to the vibration of the other pistons. From Fig. 2 it will be seen that the real part of the total impedance increases as the distance between the pistons decreases and their number and hence total area increase. When the pistons are close together, vibrate with equal amplitude, and are separated by a small fraction of a wavelength, they approximate a single piston equal in area to the combined area of the individual pistons. It may also be seen that as the separation and frequency are increased the phase of the force arising from the vibration of the other piston is retarded and may give rise to a component out of phase with the velocity of the reference piston which lowers the radiation or fluid resistance below the value the reference piston would have alone. This occurs in curve 6, Fig. 2, for values of $\omega R/c$ greater than 0.5.

ENCLOSED BACK PISTON. Same as preceding case but with small enclosure to suppress radiation from one side of a diaphragm, or to add stiffness to the vibrating system.

The fluid impedance at the external surface corresponds to the preceding case (note that $S = \pi R^2$).

If the dimensions of the enclosure are larger than the diameter of the piston the enclosure adds approximately the same effective mass per unit area, m_A, as the unlimited fluid medium. Therefore, the effective area used in calculating the total effective fluid mass is $S = 2\pi R^2$, that is, $m_f = 2\pi R^2 m_A$.

If the length of the radiated sound wave is roughly four or more times the maximum linear dimension of the enclosure, uniform adiabatic compression of the fluid occurs. The enclosure then increases the stiffness (see Section 5, article 33) per unit area of the mechanical system by an amount

$$s_A = \frac{\rho c^2 S}{V_0} \quad \text{cm dyne}^{-1}\,\text{cm}^{-2} \qquad (12)$$

(V_0 is the volume of the enclosure). The total increase in stiffness, s_f, due to the fluid, is Ss_A.

The enclosure does not approximate a constant stiffness when the length of the sound wave is less than four times the maximum linear dimension of the enclosure.

If the purpose of the enclosure is to provide a "sink" to absorb back side radiation, absorbing material is usually placed in the enclosure. This increases the effective resistance of the vibrating system. The absorption coefficient of the material used is normally high enough at the high resonant frequencies of the enclosure to make the enclosure approximate a semi-infinite medium.

3. HORNS

"A horn is an acoustic transducer consisting of a tube of varying sectional area."

The proper use of a horn as the radiating portion of a loudspeaker leads to better control of the response, efficiency, and directional characteristics. In addition these characteristics may be controlled almost independently of one another.

The total radiation response of a horn considered as an acoustical circuit element is determined largely by its throat impedance as a function of frequency. In all well-designed horns, transmission losses are minimized, so that the energy output is closely equal to the input.

The rigorous calculation of the throat impedance is possible for but very few useful horn contours, and so approximate methods are used. The low-frequency region is of greatest interest to the horn designer, as all horns have a high-frequency throat mechanical impedance which approaches a constant resistance that is the same (per unit area) for all horns. Experiment reveals that the low-frequency wave fronts in the horn are smoothly curved surfaces (Fig. 3); by expressing this mathematically there results a pressure wave equation in which the only space variable is the axial distance. For convenience, all horn-design work is usually referred to a straight axis horn of circular cross-section, as in Fig. 3.

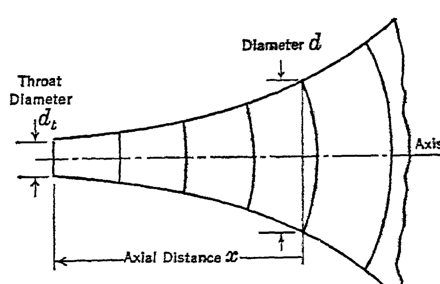


FIG. 3. Low-frequency Wave Fronts in Long Straight Axis Horn of Circular Cross-section

The question arises as to what types of horns have this kind of simplified behavior. Starting from the experimental data it is observed that the "chords" (the diameter, d, in Fig. 3) for the wave fronts are approximately proportional to the square root of the area of the wave front, provided that the expansion of the horn is not too rapid. If the sound pressure is assumed to decrease steadily as the horn expands, modified by the change of phase down the horn, this assumption may be expressed analytically and inserted into the pressure wave equation together with the relation between d and the area of wave front. This leads to a relation between d and x, the axial distance, which when solved yields

$$d = d_t[\cosh(x/x_0) + T \sinh(x/x_0)] \tag{13}$$

Here d_t is the diameter at the throat; x_0 is a reference distance fixing the rate of taper of the horn, and is related to the cutoff frequency f_c by $f_c = c/2\pi x_0$; and T is a parameter by which a particular horn contour is selected. The names "catenoidal horns" and "Salmon horns" have been suggested for this family, the latter after the person who first described their characteristics.

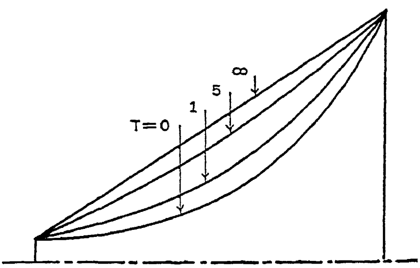
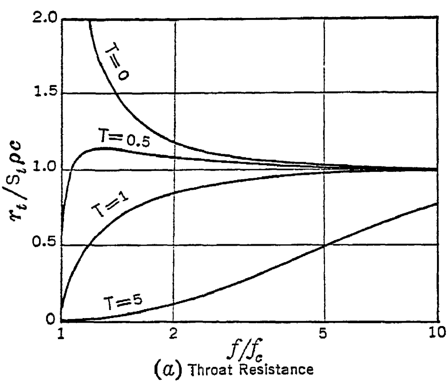


FIG. 4. Contours of Horns of Eq. (13) for $T = 0$, 1, 5, and ∞

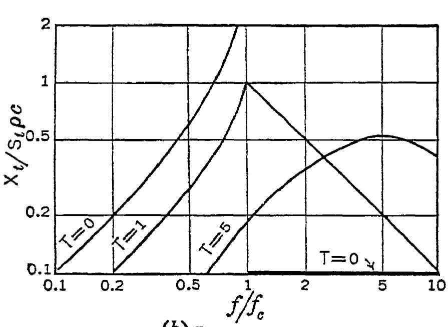
The family is the most general one consistent with the simplifying assumptions made; for $T = 1$ there results the familiar exponential horn $d = d_t \exp(x/x_0)$, while for $T = \infty$ there results, by a limiting process, the conical horn. At $T = 0$, $d = d_t \cosh(x/x_0)$, which may be termed the cosh horn. Since this family includes the most widely used horn contours, it will be taken as a basis of discussion. Contours for $T = 0$, 1, 5, and ∞ are shown in Fig. 4.

The low-frequency throat impedance of all practical horns shows considerable variation with frequency due to reflections from the mouth. Hence it is common engineering practice to state as the throat impedance that for the infinite horn (outgoing wave only); in practice the actual impedance varies about the infinite horn value as a mean, or trend, which is approached as reflections from the mouth are decreased. When the mechanical throat impedance is evaluated for the horns of eq. (13) there results

$$Z_t = r_t + jx_t$$
$$= S_t\rho c\left\{\frac{[1 - (f_c/f)^2]^{1/2} + j(Tf_c/f)}{1 - (1 - T^2)(f_c/f)^2}\right\}, \tag{14}$$



(a) Throat Resistance



(b) Throat Reactance

FIG. 5. Throat Impedance of Infinite Horns of Eq. (13). The f_c in the abscissa is the cut-off frequency.

where S_t is the throat area; ρc is the characteristic impedance of the medium; the cutoff frequency $f_c = c/2\pi x_0$, c being the velocity of a free sound wave. An examination of eq. (14) reveals that r_t is zero for $f < f_c$ (for the infinite horn), so that below f_c the impedance is entirely reactive. Figure 5 shows $r_t/S_t\rho c$ and $x_t/S_t\rho c$ as functions of f/f_c for $T = 0$, 0.5, 1, and 5. The equivalent circuit of the mechanical impedance of eq. (14) is as shown in Fig. 6; note that the parallel elements are simple; only m_t varies with T, and only r_t varies with f.

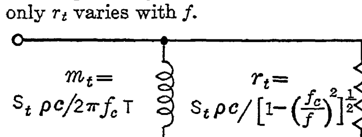


FIG. 6. Equivalent Circuit of Throat of Horns of Eq. (13). Negligible reflection from mouth.

The selection of the optimum member of the above family of horns cannot be considered apart from the loudspeaker unit or other source out of which the horn is to work, as both form a fairly tightly coupled system. The essential mechanical elements are shown in Fig. 7 for a moving-coil driving unit as discussed in article 6.

The electrical system consists of a generator of constant emf, e, and total electrical impedance, r_e, which is here assumed resistive and equal to the generator resistance proper plus the blocked resistance of the voice coil. The resulting force input to the circuit of Fig. 7 is thus $\beta le/r_e$, which is fairly constant for horn loudspeakers. As seen from the mechanical mesh, the electrical side also contributes a "source impedance" $r_1 = (\beta l)^2/r_e$ which is as constant as the force. In these expressions β is the magnetic flux density in the gap and l is the conductor length.

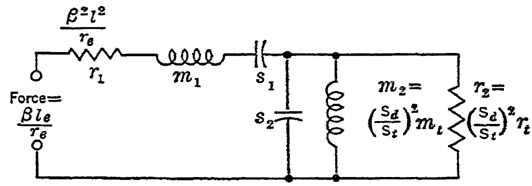


FIG. 7. Approximate Equivalent Circuit of Typical Horn Loudspeaker. $(S_d/S_t)^2$ is the impedance ratio of the fluid transformer formed by the areas S_d and S_t.

In the purely mechanical system m_1 is the mass of the loudspeaker motor, and s_1 its stiffness, which may include air trapped back of the diaphragm. The stiffness of the air s_2, of volume V, between diaphragm and horn throat (the sound chamber), is equal to $\rho c^2 S_d^2/V$ when referred to the diaphragm area S_d. Because S_d is usually different from S_t the horn throat elements of Fig. 6 appear at the diaphragm multiplied by the fluid transformer impedance ratio $(S_d/S_t)^2$, which is controlled by the horn.

The configuration of the circuit is that of a band-pass filter; however, this is not a satisfactory basis of design, as only a half section is present. In practice a flat response is not always the desired goal, and so the elements are chosen with a particular application in mind. Usually the unloaded resonant frequency, $f_1 = \frac{1}{2\pi}\left(\frac{s_1}{m_1}\right)^{1/2}$, of the motor is placed below the midband, while the other resonant frequency, $f_2 = \frac{1}{2\pi}\left(\frac{s_2}{m_2}\right)^{1/2}$, is placed above. The bandwidth is fixed largely by s_2/s_1; thus the sound-chamber volume and hence the clearance to the diaphragm should be as small as possible for a wide band. The horn influences this by controlling m_2 through the parameter T, permitting f_2 to be properly placed, with a suitable s_2.

The sound chamber, which may be regarded as a part of the horn just as a transformer is often associated with a loudspeaker, serves to reduce the effect of m_1 at high frequencies by the $m_1 - s_2$ resonance. Similarly at low frequencies the $m_2 - s_1$ resonance, as influenced by the value of T, may produce a rise in response. This reactance annulling permits good low-frequency response to be obtained by placing the horn cutoff frequency, f_c, below f_1, the resonant frequency of the $m_1 - s_1$ combination.

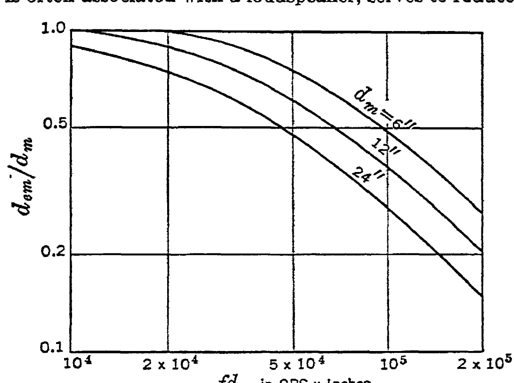


FIG. 8. Approximate Relation between Effective Mouth Diameter and Frequency for Exponential Horn

The termination r_2 is usually close to r_1 to obtain approximate matching in the mid-response region, thus governing the maximum efficiency. This involves the proper relation between horn-throat area and useful magnetic energy in the air gap. When $r_1 \doteq r_2$, variations in horn-throat impedance due to mouth reflections have a minimum effect on the radiated power. These reflections are minimized also when the product of mouth diameter in inches and cutoff frequency in cps is greater than 4000; this product may be made as little as 2000 if $r_2 \doteq r_1$.

Since diaphragms are often called on to radiate energy at wavelengths for which destructive interference may take place across the diaphragm, it is usual to remove the radiation by one or more annular slots so placed that phase effects are minimized. The annular passages may then be constricted in average diameter until the circular horn section is reached.

The directional properties of a horn are largely determined by the mouth geometry, particularly the diameter and slope. As frequency is increased, there is reached a value at which the emergent sound is so directional that the wave does not "touch" the mouth portion at all. Thus the effective mouth diameter decreases as the frequency is raised, and in some designs the polar response pattern becomes almost independent of frequency. Data reported in the literature may be presented in the form of equivalent mouth diameter as a function of frequency. In Fig. 8 this is shown for one group of measurements on exponential horns; the curves are rough approximations and should be taken as indicating only the trends. The ratio (d_{em}/d_m) is that of equivalent and actual mouth diameters.

In general, the steeper the mouth slope, the less directional the horn; but, if this generalization is carried too far, the violence done to the wave fronts appears as a rough response characteristic. The directional effect of horns is not too serious in practical applications, for often this property is useful in reducing acoustic feedback and improving the sound level in a desired localized region. When uniform radiation over a large solid angle is required, multicell horns are commonly used. Often the shape of the horn may be altered so as to provide improved distribution, and this frequently results in a more uniform space-response variation.

BIBLIOGRAPHY

American Standard Z24.1-1942. (Terminology.)
Ballantine, Stuart, *J. Franklin Inst.*, Vol. 203, 85 (1927). (Bessel horns.)
Crandall, I. B., *Theory of Vibrating Systems and Sound*, p. 85 ff. (Impedance of medium and of horns.)
Hall, W. M., *J.A.S.A.*, Vol. 3, 552 (1932). (Sound field in horns.)
Hanna, C. R., and J. Slepian, *Trans. A.I.E.E.*, Vol. 43, 393 (1924). (Horn loudspeakers.)
Hoersch, V. A., *Phys. Rev.*, Vol. 25, 218 and 225 (1925). (Transverse vibrations in horns.)
International Critical Tables, Vol. 4, 453 (1929). (Constants.)
I.R.E. Standards on Electroacoustics (1938). (Terminology.)
Klapman, S. J., *J.A.S.A.*, Vol. 11, 289 (1940). (Multiple pistons.)
Lamb, H., *Proc. Roy. Soc. London*, Vol. 98, 205 (1920). (Fluid impedance.)
McLachlan, N. W., *Loudspeakers* (1934). (Fluid impedance, general, and bibliography.)
McLachlan, N. W., and S. Goldstein, *J.A.S.A.*, Vol. 6, 275 (1935). (Finite pressure distortion horns.)
Mason, W. P., *B.S.T.J.*, Vol. 6, 258 (1927). (Horns and other acoustic elements.)
Phelps, W. D., *J.A.S.A.*, Vol. 12, 68 (1940). (Finite horn boundary impedance.)
Rayleigh, J. W. S., *Theory of Sound*, Vol. II, 164. (Fluid impedance.)
Salmon, V., *J.A.S.A.*, Vol. 17, 199 and 212 (1946). (Horns.)
Stewart, G. W., and Lindsay, R. B., *Acoustics*, p. 132. (Horns and general.)
Thuras, A. L., R. T. Jenkins, and H. T. O'Neil, *B.S.T.J.*, Vol. 14, 159 (1935). (Finite pressure distortion horns.)
Van Urk, T., and R. Vermeulen, *Philips Tech. Rev.*, Vol. 4, 213 (1939). (Radiation admittance.)
Webster, A. G., *Proc. Natl. Acad. Sci.*, Vol. 5, 275 (1919). (Horns.)

LOUDSPEAKERS AND TELEPHONE RECEIVERS

By Hugh S. Knowles

A loudspeaker is an electroacoustic transducer actuated by energy from an electrical system and radiating energy into an acoustical system, the spectral composition of the energy in the two systems being substantially equivalent. Loudspeakers may be classified as to type of radiator or radiating system, type of motor, and reversibility.

4. ACOUSTIC RADIATORS

DIAPHRAGMS. The transformation of electrical into acoustical energy is usually accomplished by electrically actuating a surface or diaphragm in contact with air, or some other fluid, causing it to move and set the adjacent air particles in motion. (See, however, article 9.) When the resulting radiation is into a large solid angle the radiation resistance, or real part of the fluid impedance, is low when the length of the radiated wave substantially exceeds the diameter of the radiator. The diaphragm serves to couple the air, which has low impedance per unit area, to a motor having a relatively high mechanical impedance, when connected to its source of electrical energy, viewed from the diaphragm.

The low radiation resistance is unfortunate both because of the problem of obtaining efficient energy transfer and because large diaphragm amplitudes are required to radiate appreciable power at low frequencies. Figure 1 shows the peak diaphragm amplitude, or half the total diaphragm displacement, required of several piston sizes to radiate 1 watt. One-tenth this amplitude is required to radiate 10 milliwatts. These curves cover the frequency range in which the radiation resistance is proportional to the square of the frequency and to the fourth power of the piston radius.

Most diaphragms are conical in shape and, if rigid, displace the same amount of air for the same amplitude as would a flat piston having a diameter equal to the diameter of the base of the cone. At low frequencies the base diameter of the cone is therefore taken as the equivalent piston diameter.

The size of the usual conical diaphragm, frequently called a cone, is limited by the need for increasing its mass per unit area as the area is increased in order to maintain adequate stiffness. It is customary to make the diaphragm thickness vary almost directly with the diaphragm diameter. This largely offsets the improved resistance-reactance ratio of the fluid impedance.

The conical diaphragm may be thought of as a conical transmission surface in which there are both radial and circumferential waves. At low frequencies only the radial wave need be considered. The flexible annular support or "surround" provides a termination for the base or large diameter of the conical surface. The flexural phase velocity of the radial wave in the diaphragm and the impedance of the termination are such that the effective length is one-quarter wave at 600 to 1000 cps in large diaphragms (16-in. to 10-in. "pistons") and at 1000 to 2000 cps in small diaphragms (8-in. to 2-in. "pistons"). Below this frequency all parts of the cone move in phase although the amplitude is not substantially uniform except at lower frequencies where the radial length is a small fraction of a wavelength. If the annular support behaves as a non-linear stiffness over the required amplitude range, as it frequently does, the diaphragm may flex in a complicated way even at low frequencies.

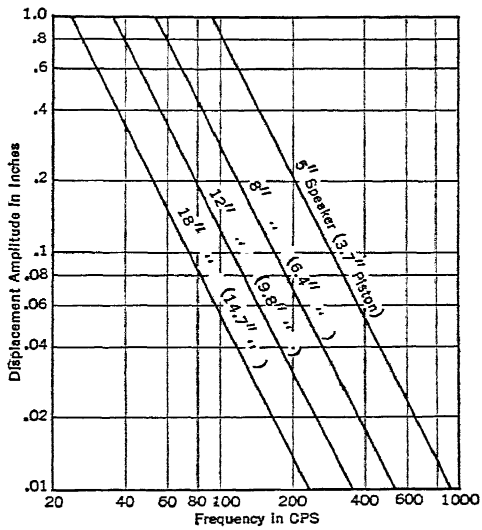


FIG. 1. Displacement of Piston from Equilibrium Position Required to Radiate 1 Watt. The total displacement is twice the value given. (If both sides radiate, divide by $\sqrt{2}$.)

Wave transmission in the diaphragm is desirable because it results in a more favorable high-frequency driving point impedance into which the motor is coupled. It also results in a broader high-frequency directional pattern than would obtain if the diaphragm were a rigid piston. When maximum loudness is required, a diaphragm made of materials having low internal dissipation such as pressed or calendered paper and a low-resistance flexible annular termination is used. When a smoother response-frequency curve is desired with reduced transient distortion and the reduced loudness can be tolerated, a "soft" more blotterlike material with higher flexural resistance is employed, sometimes with a dissipative termination of leather, felt, cloth, or a dissipative elastomer. The modes of vibration of the diaphragm are also influenced by cone angle, lumped masses and compliances (usually annular beads or corrugations), impregnants, etc. Because of the complex behavior of diaphragms most design work is largely empirical.

To reduce the mass per unit area and yet obtain the benefit of a large radiating area multiple diaphragms are frequently used. If all the diaphragms vibrate in phase and with the same amplitude the average radiation impedance seen by the array will correspond to that of a single diaphragm of equal area. (See p. 5-66.) By properly orienting the speakers the spatial radiation pattern, at high frequencies, may be improved.

In addition to the common right circular cone shape two others are sometimes used. The conoidal or "curvilinear" is used when an increase in radiation above 5000 cps is wanted at the expense of the 2000–5000 cycle region. The elliptical shape is sometimes used when space requirements limit one dimension of the cone. Unfortunately the limited dimension is usually the vertical one, resulting in the major or long axis of the ellipse being mounted horizontally. Contrary to popular belief this leads to a narrow horizontal and wide vertical high-frequency directional pattern as predicted from theoretical considerations. In spite of the appeal of its shape the oval or elliptical diaphragm has had limited acceptance because it is more difficult to fabricate, its response is more difficult to adjust, it has less radiation resistance in the middle-frequency range than a circular diaphragm of identical area, and its asymmetrical support leads to cone and moving-coil distortion which necessitate larger air-gap clearances for a comparable safety factor.

Over the frequency range in which the fluid mass, m_f, is substantially constant most conical diaphragms behave as a rigid piston of mass m_m. If the flexible centering members supporting each end of a conical diaphragm are linear (displacement proportional to force), their stiffnesses may be combined into an equivalent stiffness s_m or equivalent compliance C_m.

HORNS. The radiation from a horn may be considered as originating at the mouth, with the remainder of the horn serving to "match" the mouth and throat terminating impedances. For this purpose it is important that the walls be relatively non-porous and non-vibratile.

The theory has been developed on the basis of a straight axis horn of circular cross-section, with no reflections from the mouth. One common departure from these idealized conditions is in the shape of the axis. It may be bent, as in the low-frequency horns used in theater systems, or it may be folded, as in re-entrant public-address horns. The contour of the boundary at the changes in the direction of the axis must usually be selected empirically because of the complex manner in which the wave front executes the change in direction. The contour must also be selected so as to avoid exciting transverse modes of vibration of the air. This requires smoothly and symmetrically changing contours, and careful attention to the symmetry of the driving diaphragms with respect to the throat of the horn.

In practice, the cross-sectional area may be circular, rectangular, or annular, the last two corresponding to the theater and public-address horns mentioned above. In these horns the lateral dimensions are chosen to retain the same area-axial distance relation as in the circular (reference) horn.

The effect of mouth reflections may be minimized by proper attention to the horn driver unit (see article 3). Advantage may be taken of reflections to load the horn unit at frequencies near cutoff for which the radiation resistance is normally low.

5. EFFICIENCY

The total power dissipated by the mechanical system is

$$P_{mf} = \dot{s}^2 r_{mf} = \dot{s}^2 (r_m + r_f) = \frac{I^2 z_{12}^2}{z_{22}^2}(r_m + r_f) \qquad (1)$$

in which r_m and r_f are the mechanical and fluid resistances, I is the current in the electric mesh and z_{22} is the total self-impedance of the mechanical system, and the electrical and mechanical meshes are numbered 1 and 2 respectively. (See Section 5, article 33 and eqs. [4*a*] to 5[*b*] below.)

Energy Efficiency. The ratio of acoustic power (or energy) output to electric power (or energy) input is called the energy or conversion efficiency and corresponds to the definition of efficiency commonly used for most transducers other than electroacoustic ones. This efficiency is

$$\eta_e = \frac{z_{12}^2 r_f}{z_{22}^2 r_{mc} + z_{12}^2 r_f} \qquad (2)$$

where r_{mc} is the blocked resistance of the system.

System Efficiency. Because both the modulus and phase angle of the normal input impedance of the speaker vary with frequency, the speaker in general absorbs less power from the source than an ideal load or transducer would. Since this inability of the transducer to absorb maximum power limits the useful power output of the source, it in effect reduces the "efficiency" of the transducer.

To take this property into account, the ratio of the acoustic power output to the electric power input which the source would supply if connected to an ideal transducer is defined as the system or absolute efficiency. If the source is a vacuum tube or generator whose internal impedance is a pure resistance, r_g, it supplies maximum power to a resistance of equal value, and the system efficiency is

$$\eta_s = \left| \frac{4 r_g z_{12}^2 r_f}{(z_{11} + z_M)^2 z_{22}^2} \right| \qquad (3)$$

where $z_M = z_{12}^2/z_{22}$, the other quantities are as defined above, and the vertical lines indicate that the absolute value is to be taken.

6. MOVING-CONDUCTOR SPEAKERS

"A *magnetic speaker* is a loud speaker in which the mechanical forces result from magnetic reactions."

"A *moving-conductor speaker* is a magnetic speaker in which the mechanical forces result from magnetic reactions between the field of the moving conductor and the steady applied field. (This is sometimes called a dynamic speaker.)" This classification includes moving-coil or electrodynamic (sometimes called "dynamic") and ribbon or "band" speakers. Cross-sections of two typical moving-coil speakers are shown in Fig. 2.

The mechanical circuit may be considered a series circuit with the elements m_{mf}, C_{mf}, and r_{mf}, which include the fluid impedance. In general the circuit elements are functions of frequency.

The actuating force on the conductor is βli dynes, where β is the flux density in gausses, l the conductor length in centimeters, and i the instantaneous current in abamperes (amperes × 10). If the electrical circuit, including the moving coil and generator, has a total series inductance L_e, capacitance C_e, and resistance r_e, and the generator an instantaneous open-circuit voltage e, the instantaneous "force" equations are

$$L_e\ddot{q} + r_e\dot{q} + \frac{q}{C_e} + \beta l\dot{s} = e \qquad (4a)$$

$$m_{mf}\ddot{s} + r_{mf}\dot{s} + \frac{s}{C_{mf}} - \beta l\dot{q} = 0 \qquad (4b)$$

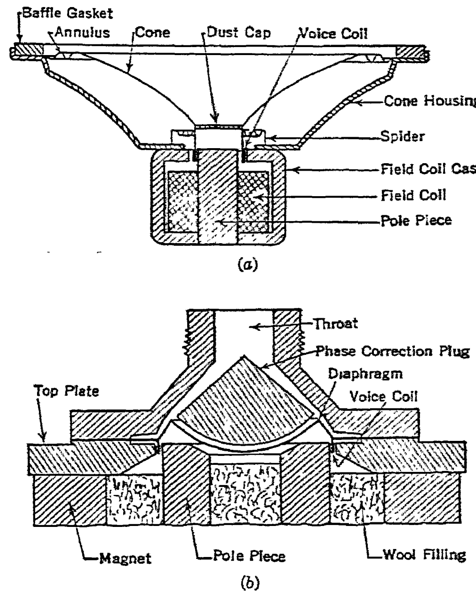


FIG. 2. Moving-coil Speakers. (a) Direct radiator type; (b) horn type.

The further assumption is made in these equations that β is a constant, independent of s, that $\dot{q}$ does not alter the magnetic field, and that there is negligible mutual impedance between the moving coil and the "field" coil which provides the static field. If β is supplied by a permanent magnet, a constant flux source is approximated which is little altered by $\dot{q}$. The performance of moving-coil speakers depends on β and not on whether this is supplied by a permanent magnet or an electromagnet.

If $e = E \sin \omega t$, the steady-state solution of eqs. (4) is

$$E = z_{12}\dot{s} + z_{11}I \qquad (5a)$$

$$0 = z_{22}\dot{s} - z_{12}I \qquad (5b)$$

where E, I, and $\dot{s}$ are the rms steady-state voltage, current, and velocity, and $z_{12} = \beta l$ is the force factor. "The force factor of an electroacoustic transducer is a measure of the coupling between its electrical and mechanical systems. It is the ratio of the open-circuit force or voltage in the secondary system to the current or velocity in the primary system." Equations (5) give

$$I = \frac{Ez_{22}}{z_{11}z_{22} + z_{12}^2} \qquad (6a)$$

$$\dot{s} = \frac{Ez_{12}}{z_{11}z_{22} + z_{12}^2} \qquad (6b)$$

DIRECT RADIATOR OR HORNLESS SPEAKER. These are the commonly used electrodynamic speakers which are designed to radiate as efficiently as possible into a solid angle of the order of 2π("semi-infinite medium"). They are used in baffles or cabinets when response at low frequencies is required but space limitations prevent the use of a horn or "directional baffle" to improve the efficiency. The radiated sound energy is $\dot{s}^2 r_f$, where $\dot{s}$ is given in eq. (6b) and r_f is Sr_A, with r_A from Fig. 1, p. 13-03, provided the mounting approximates an infinite baffle at the frequency considered. In the low and low middle frequencies, the fluid resistance is proportional to the square of the frequency. Therefore, if $\dot{s}^2 r_f$ is to be approximately constant, the velocity $\dot{s}$ must vary inversely with frequency. From eq. (6b) it may be seen that, when the applied frequency is appreciably higher than the resonant frequency of the mechanical system, $\dot{s}$ is largely a function of z_{12} and z_{22}. Since z_{12} is approximately a constant, and $z_{22} \doteq \omega m_{mf}$, $\dot{s}$ varies inversely with ω particularly if z_{12} is small. The radiated acoustic power is therefore approximately independent of frequency. In this frequency range the system efficiency, η_s, is approximately

$$\eta_s \doteq \frac{2 \times 10^{-9} \alpha E C}{(m/d^2)^2} \tag{7}$$

where α is the ratio of actual conductor volume to air-gap volume, V (product of gap length, mean perimeter, and coil winding length);

$$E = \frac{\bar{B}^2 V}{8\pi} \doteq \text{gap energy (ergs)}$$

$\bar{B}$ is the average flux density in volume V (gauss); C is the conductivity of the conductor with respect to copper; m is the effective motor mass (voice coil, cone, fluid mass) (gram); and d is the effective piston diameter (inches).

When the velocity of the diaphragm is largely limited by the mass reactance of the mechanical system and varies (roughly) inversely with frequency, the device is said to have "inertia control." To obtain uniform response, the natural frequency is placed near the lowest frequency to be transmitted. In 12-in. speakers this resonant frequency is of the order of 75 cps. At very low frequencies, either the stiffness reactance or the mechanical resistance limits the velocity, and the response decreases even if the speaker is operated in an infinite baffle. In the middle frequency range, the normal impedance and, therefore, the force on the mechanical circuit are fairly constant. In general the damping is less than "critical" or the value required to give the shortest transient response. See "Transient Response," Section 5, articles 13 and 15.

Figure 3 indicates the result of varying βl (or z_{12}) on the pressure measured in the sound field. Since the speaker directivity is almost independent of frequency in this range, the ordinates are proportional to the total radiated acoustic power. These curves show that

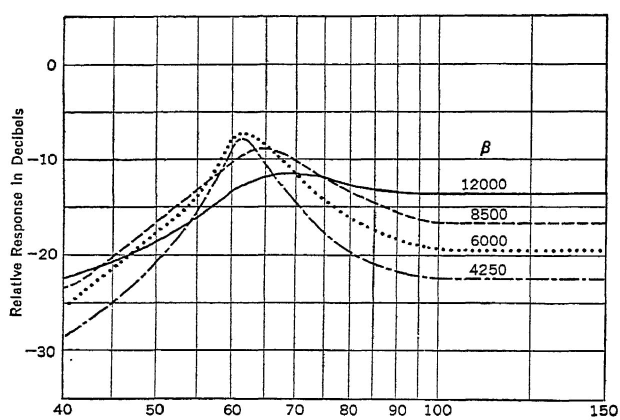


FIG. 3. Variation in Acoustic Pressure with Flux Density, or Force Factor. (Measurements were made on the axis, 10 feet from a direct radiator moving conductor speaker.)

speakers having large force factors or values of z_{12} and high efficiency in the middle frequency range have relatively less response near the resonant frequency of the mechanical circuit. Increasing the flux density and force factor increases the energy efficiency but decreases the system efficiency near resonance. Because the maximum peak powers in speech and music (see Section 12, articles 10 and 11) occur in the middle frequency range, it is important that the speaker have high efficiency in this range, and the better speakers have high force factors, which also reduce the transient distortion of the speaker.

Direct radiator speakers normally reach their maximum efficiency in the low middle frequency range, except when the effective baffle size is large so that the resonant frequency is radiated effectively. In that case, if z_{12} is small the maximum system efficiency may occur at the frequency of mechanical resonance. The normal impedance is resistive, or largely so, in either case. (See Fig. 4.)

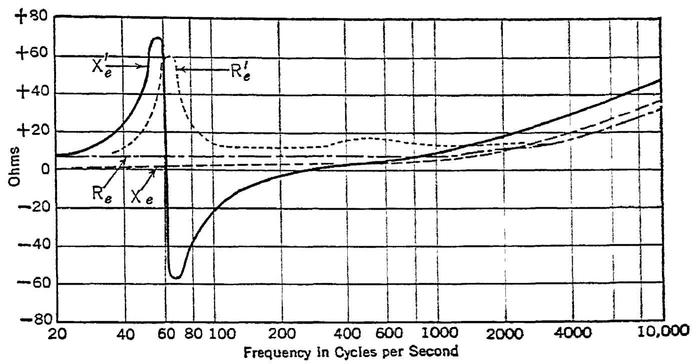


FIG. 4. Blocked and Normal (Primed) Resistance and Reactance of Moving-coil Direct-radiator Speaker in Ohms

The system efficiencies in the middle frequency range vary from roughly 1 to 20 per cent in commercial direct radiator speakers. The energy efficiency values are slightly higher in the middle frequency range and exceed 90 per cent at the mechanical resonant frequency in some designs.

ENCLOSURES. Direct radiator speakers are often placed in enclosures to contain and control the radiation from the rear surface of the diaphragm. If the enclosure is complete, that is, has no vents, it is sometimes called a "total enclosure" (see Enclosed Back Piston, article 1). The rear radiation may be put to use by permitting it to escape from the enclosure after first modifying it in phase and magnitude by a suitable acoustic network. When the vents, or ports, which radiate the energy from the rear of the diaphragm are close to the front surface of the diaphragm, the mutual impedance is high. In this case the phase of the rear radiation is critical but gives rise to maximum radiation when it is properly controlled (see article 2).

When the length of the radiated wave is large compared with the enclosure dimensions the acoustic parameters may be considered "lumped." A single cavity enclosure with a port may then be considered to add a compliance corresponding to that seen by the diaphragm with the port blocked or covered, which is in series with the speaker compliance. The enclosure compliance is shunted by the effective mass and radiation resistance of the port. The latter values are those referred to the speaker diaphragm. The frequency of the two resulting low-frequency modes may be computed very approximately by neglecting the mutual reactance term arising from the coupling of the external radiating surfaces of the diaphragm and port through the mutual fluid impedance. The real part of this mutual impedance must, however, be included in any accurate calculation of the total radiated power.

The vented enclosure may be used to maintain uniform radiation down to lower frequencies than those radiated effectively by a total enclosure of identical volume, or it may be used to provide a rise in the radiation near the cutoff frequency of the unvented enclosure. In practice it is common to provide a compromise between these two by so choosing the speaker and enclosure parameters that the resulting two modes of the simple structure described above occur approximately one-half octave below and above the mode of the speaker and enclosure with blocked port. The augmented low-frequency radiation is obtained with substantially reduced non-linear distortion since most of the radiation is from the port, which need have no variable or non-linear parameters, and the diaphragm displacement is appreciably reduced.

The rear of the diaphragm is sometimes coupled to a highly absorbent line approximately one-quarter wave long near the cutoff frequency of the acoustical system. The line absorption should be low in the frequency interval in which the line is between three-quarters and one wave long and increase rapidly above the upper frequency of this interval to suppress radiation which would otherwise be out of phase with the front radiation. When the line is one-quarter and three-quarter wave long it serves as an impedance inverter, thereby raising the impedance seen by the diaphragm and reducing its amplitude for a given total radiation.

HORN-TYPE MOVING-COIL (OR CONDUCTOR) SPEAKERS. To increase the fluid resistance, r_f, a horn may be used to couple the diaphragm to the acoustic medium. Because of their greater throat resistance at low frequencies for a given size, hyperbolic-exponential horns with $T \leq 1$ (see article 3) are normally used.

In the case of horn units the normal (electrical) impedance can be made quite uniform (see Fig. 5) so that the force on the diaphragm is approximately constant if the source voltage and impedance are constant.

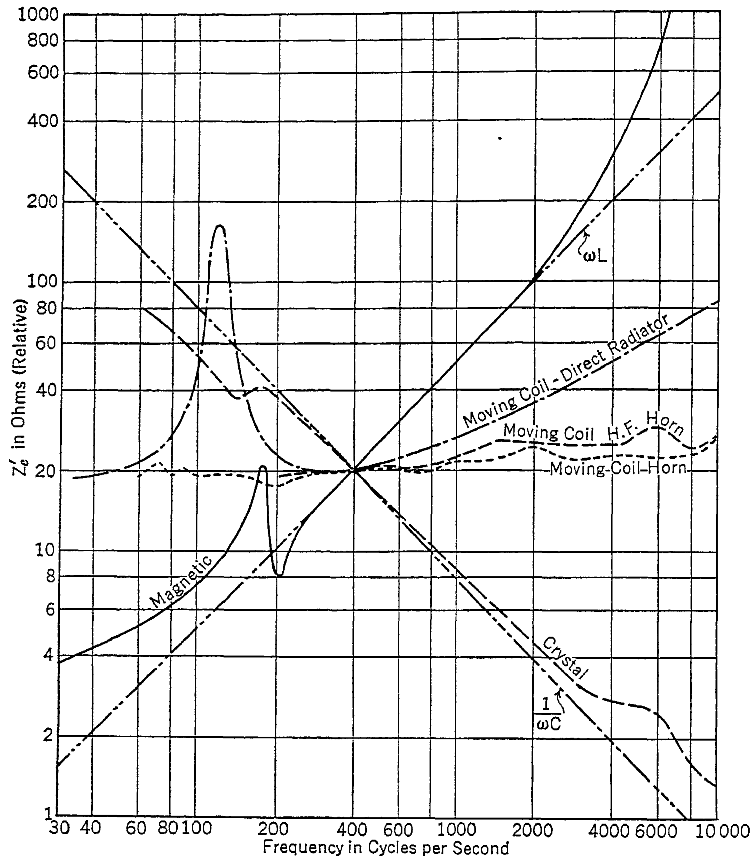


FIG. 5. Relative Scalar Normal Impedances of Typical Loud Speakers. (All arbitrarily adjusted to have the same impedance at 400 cps. Reactance of an inductance and capacitance plotted for comparison.)

The diaphragm is made to resonate an octave or two above the horn cutoff frequency. The stiffness reactance of the diaphragm assembly then reduces the effect of the mass reactance of the horn near its cutoff frequency. At high frequencies, the stiffness of the sound chamber may be made to reduce the effect of the mass reactance of the diaphragm. The fluid resistance is adjusted to the desired value by proper choice of the ratio (S_d/S_t). These factors make possible the design of a horn having uniform response over an extended frequency range.

At low frequencies, irregularities in response result from the variation in throat impedance of a horn of finite length. The resonant frequencies of the horn near its cutoff frequency also give rise to transient distortion. By using a horn unit having a large force factor, and by proper choice of the other elements, the steady-state variation in response and the transient distortion may be made negligible.

At high frequencies destructive interference occurs in the sound chamber since its dimensions are of the order of magnitude of the length of the sound wave. The horn throat is therefore sometimes so constructed that the difference in path length from different parts of the diaphragm to the throat is a minimum. The use of a phase equalizing plug for this purpose is shown in the horn unit of Fig. 2. At very high frequencies the diaphragm no longer behaves as a rigid piston, and this alters the response of the unit.

The minimum compliance of the sound chamber is limited by its minimum volume, V_0. This volume must be adequate to provide ample mechanical clearance at low frequencies, where the diaphragm amplitude is a maximum, and to limit the distortion which results from "finite" sound pressures.

7. MAGNETIC-ARMATURE SPEAKERS

"A magnetic-armature speaker is a magnetic speaker whose operation involves the vibration of the ferromagnetic circuit. (This is sometimes called an electromagnetic speaker.)" Numerous magnetic-armature designs have been proposed. The principle of operation of this type of speaker is analogous to that of moving-conductor speakers, and equations for the latter apply to this type if the appropriate value of z_{12} is used. They differ principally in the fact that the conductor does not move, which permits the use of a large conductor volume. In practice, high loudness efficiencies can be obtained in a limited frequency range with a magnet having moderate magnetomotive force. The efficiency at very high frequencies is normally poor because of losses in the armature. The very low-frequency response is limited by the stiffness required to give adequate armature stability. The current displacement curve is linear over a very limited range and gives rise to appreciable amplitude distortion.

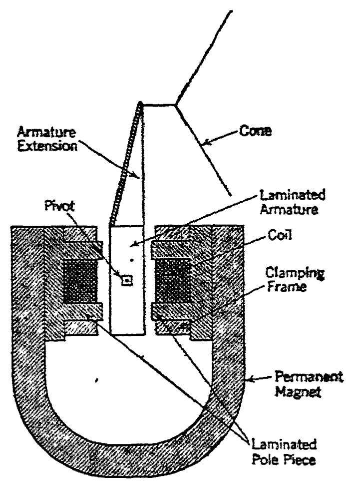


FIG. 6. Balanced Armature Speaker

BALANCED-ARMATURE MAGNETIC SPEAKER. A cross-sectional view of a unit of this type is shown in Fig. 6. The speech current flows through the stationary coil which surrounds the armature. It increases the effective flux through two (diagonally located) gaps and decreases it in the other two gaps. The steady flux, ϕ_0, is increased and decreased by an amount $\phi = 4\pi Ni/\mathcal{R}$ gausses, where N is the number of turns on the coil, i is the current through the coil in abamperes, and $\mathcal{R}$ is the effective reluctance of the alternating flux path.

The total effective force on the armature, acting at the magnetic force center, is

$$F = \frac{2(\phi_0 + \phi)^2}{8\pi A} - \frac{2(\phi_0 - \phi)^2}{8\pi A} = \frac{4\phi_0 NI}{\mathcal{R}A} = \frac{4\beta_0 NI}{\mathcal{R}} \tag{8}$$

where F is the rms force in dynes and A is the effective area in square centimeters at each gap and β_0 is the steady or undisturbed flux density. For small amplitudes the reluctance of the alternating flux path is approximately constant, since most of the reluctance is in the gap, and the reluctance of the permanent magnet path is large.

The assumption is usually made that all the elements of the circuits and the force factor are constant. This is approximated only when the amplitude is small. When the armature is in the undisturbed center position, it is in a state of unstable equilibrium. The application of a force results in a force tending to increase the displacement. This property is called negative stiffness or compliance. (See Section 5, article 30.) It here results from the fact that the torque on the armature is proportional to the square of the flux density. Therefore, it increases more rapidly at the tip that is approaching one magnet tip than it decreases at the tip that is receding from the other magnet tip. Since the force varies with the square of the flux density, and the density varies almost inversely with distance between the armature and magnet tip, the stiffness can be considered a constant only for small displacements. The positive value of C_m is, therefore, the difference between the positive diaphragm (and fluid, if any) compliance and the negative compliance. Stability of the system requires that the rate of change of force with displacement be less than the coefficient of stiffness. The further assumption is made that the lever arm has infinite stiffness. In practice, resonances of this arm play a large part in modifying the mechanical impedance at high frequencies.

The steady-state solution for this speaker is the same as that of eqs. (5) and (6). In this case $z_{12} = 4\beta_0 N/\mathcal{R}$; for the more exact theory z_{12} is considered complex.

The principle of operation is analogous to that of the moving-conductor speaker. Equations for the latter apply if the above value of z_{12} is substituted. A large conductor volume is normally used to increase the force factor. This makes the inductance of the electrical circuit much higher than in moving-coil speakers. For this reason, the normal impedance varies over wide limits. The normal impedance is proportional to frequency over much of the frequency range, as shown in Fig. 5. The rise in impedance above the value predicted on this basis at the high-frequency end, in this particular speaker, was due to electrical resonance. The speaker was of the high-impedance type, and the inductance and distributed capacitance resonated at the high-frequency end. This is normally the case in high-impedance speakers of this type, and its effect must be included in any complete performance analysis.

BIPOLAR MAGNETIC-ARMATURE SPEAKER. This type of speaker consists of a steel diaphragm mounted near the two ends of a U magnet. The speech current flows through the stationary coils which surround the two pole pieces. If the current through the two coils, connected in series, is $I \sin \omega t$, and the coils have N turns, the alternating flux is $\phi = \dfrac{4\pi NI \sin \omega t}{\mathcal{R}}$, where I is the maximum value of the current in abamperes, and $\mathcal{R}$ is the effective reluctance of the alternating flux path.

The total effective force F acting at the magnetic force center is

$$F = \frac{2(\phi_0 + \phi)^2}{8\pi A} = \frac{\phi_0^2}{4\pi A} + \frac{2\phi_0 NI \sin \omega t}{\mathcal{R}A} + \frac{2\pi N^2 I^2}{\mathcal{R}^2 A} - \frac{2\pi N^2 I^2 \cos 2\omega t}{\mathcal{R}^2 A} \tag{9}$$

The first term on the right is a constant force; the second is the useful component which is proportional to ϕ_0 and to the signal current. The last two terms are distortion terms and are minimized by making $\phi_0 \gg \phi$.

The force factor $z_{12} = 2\beta_0 N/\mathcal{R}$. The more exact theory includes the case in which z_{12} is complex.

The principle of operation is analogous to that of the balanced-armature magnetic and moving-conductor speakers. The equations for these speakers apply to the bipolar magnetic armature if the value of z_{12} given above is used.

8. CONDENSER SPEAKERS

"A condenser speaker is a loud speaker in which the mechanical forces result from electrostatic reactions." Speakers of this type have a movable conducting electrode which serves as the diaphragm and is mounted close to a perforated fixed electrode or between two perforated fixed electrodes.

TWO-ELECTRODE CONDENSER LOUDSPEAKER WITH MECHANICAL CIRCUIT HAVING ONE DEGREE OF FREEDOM. One speaker of this type has a movable and a fixed electrode separated by a thin dielectric as shown in Fig. 7. The force of attraction between the electrodes as the charge varies provides the actuating force for the movable electrode which is used as the diaphragm.

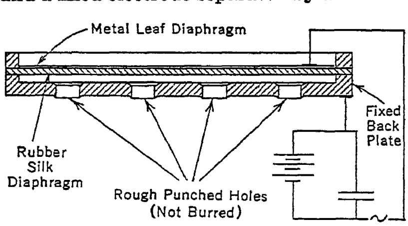


FIG. 7. Condenser Speaker

The equations for a system of this type are developed in Section 5, article 32. The steady-state equations for a single sine-wave applied voltage are:

$$E = z_{11} I + z_{12} \dot{s} \tag{10a}$$

$$0 = z_{12} I + z_{22} \dot{s} \tag{10b}$$

where $z_{12} = \dfrac{jq_0}{\omega C_0 d_0} = \dfrac{jE_0}{\omega d_0}$. The force factor therefore varies directly with the polarizing voltage, E_0, and inversely with the frequency and no-signal separation of the electrodes, d_0. The minimum value of the separation is determined by the maximum low-frequency amplitude, and it must be large in comparison with the amplitude if non-linear distortion is to be avoided.

$$I = \frac{Ez_{22}}{z_{11}z_{22} - z_{12}^2} \tag{11a}$$

$$\dot{s} = \frac{-Ez_{12}}{z_{11}z_{22} - z_{12}^2} \tag{11b}$$

The radiated acoustic power is $\dot{s}^2 r_f \times 10^{-7}$ watt, where $\dot{s}$ is given by eq. (11b) and r_f by article 2. The diaphragm does not move as a piston. Therefore, a generalized velocity based on the type of deformation here obtained must be used.

Because the blocked impedance of the speaker is that of a capacitance, C_0, the normal impedance varies inversely with frequency over much of the frequency range. Therefore, uniform response and high efficiency are difficult to obtain. The impedance of the electrical circuit is sometimes altered to improve the response.

9. PNEUMATIC SPEAKERS

"A pneumatic speaker is a loudspeaker in which the acoustic output results from variations of an air stream." It is of the irreversible or relay type. The definition of efficiency for reversible speakers is not used in this case, since the efficiency, as defined for them, can exceed 100 per cent because no account is taken of the power used to compress the air. Any type of loudspeaker motor may be used to drive a very light balanced valve which modulates an air stream.

These units permit the generation of large acoustic powers. The valve and valve parts give rise to spurious noises that are difficult to eliminate. When large acoustic outputs are generated, there is appreciable "finite" amplitude distortion in the types that have been made.

10. TELEPHONE RECEIVERS (EARPHONES)

Telephone receivers are so constructed that they operate directly into the ear cavity. The fluid impedance of the ear cavity varies appreciably from ear to ear, particularly at high frequencies. If there is no fluid leak between the ear "cap" of the receiver, which held in contact with the ear, and the ear, the fluid impedance is approximately that of a 6-cm^3 cavity at frequencies in the middle and lower frequency ranges. At high frequencies, the ear cavity resonates and the fluid resistance of the cavity increases. If there is a fluid leak between the ear cap and ear, the fluid resistance of this leak must be considered at low frequencies where it tends to reduce the pressure in the cavity.

It has been found that a good compromise design for most ears is one in which the diaphragm displacement is independent of frequency. At low frequencies this gives constant pressure in the ear cavity if there is no fluid leak. Since the magnitude of the reactance of the ear impedance substantially exceeds the resistance below 1000 cps, the power supplied to the ear and the efficiency, as defined for other transducers, are not satisfactory performance criteria. The pressure squared, produced at the end of a non-dissipative cavity or "coupler," per unit power available from an ideally terminated source, is a common performance criterion. A 6-cc coupler is used for ear cap or external earphones and a 2-cc coupler for insert receivers which are placed in the outer ear canal and thereby reduce the volume of the cavity between the receiver and the eardrum.

MOVING-CONDUCTOR TELEPHONE RECEIVERS. Both moving-coil and ribbon telephone receivers are in use. The construction of one type of moving-coil receiver is similar to that of the microphone shown in Fig. 3, p. 13-24. The values of the circuit elements differ from those used in the microphone, in order to give as nearly as possible constant diaphragm amplitude at all frequencies with a constant-voltage source.

If the mechanical circuit consists only of the stiffness and resistance of the diaphragm, and the combined mass of the diaphragm and moving coil, its "force" equations are given by eqs. (6). The displacement $s = \dot{s}/j\omega$, where $\dot{s}$ is given by eqs. (6). Since z_{12} is a constant, if z_{11} is made approximately constant, z_{22} must vary inversely with frequency for the diaphragm displacement to be independent of frequency. This requires that the diaphragm have stiffness reactance over the range of uniform response. This is obtained by placing the resonant frequency of the diaphragm near the maximum frequency to be transmitted. Under these conditions the normal impedance of the receiver is fairly uniform below the frequencies near the resonant frequency of the diaphragm. The force is therefore nearly constant. The displacement of the diaphragm is small, and hence the efficiency of the receiver is poor when the resonant frequency and stiffness are high. It is also difficult in practice to obtain very high resonant frequencies.

To improve the efficiency and yet maintain uniform response, additional mechanical circuits are sometimes coupled to the diaphragm. The resonant frequency of the diaphragm itself is lowered, and the value of the mechanical elements is chosen to give uniform diaphragm displacement.

If the mechanical impedance is large in comparison with the ear cavity impedance, the latter may be neglected. In general, it may be neglected only for approximate calculations. The impedance of the fluid between the diaphragm and the ear cap must also be considered in any complete analysis.

MAGNETIC-ARMATURE TELEPHONE RECEIVER. The theory of operation corresponds to that of magnetic-armature loudspeakers given above. The first center moving mode of vibration of the magnetic diaphragm in the early type of telephone receiver occurs near 1000 cps. Near this frequency, the amplitude is large and the pressure in the ear cavity is large. Below this resonance region, the diaphragm has approximately constant stiffness or compliance, and the amplitude is fairly uniform. At frequencies above the resonant region, the displacement decreases and goes through a series of decreasing maxima at higher modes of vibration. In more recent types there is a thin film of air at the back of the diaphragm which is coupled to an auxiliary acoustical network. The parameters are chosen to give uniform pressure at the bottom of the test cavity or coupler up to 3000 cps in telephone hand sets and up to 4000 cps in some military types.

PIEZOELECTRIC TELEPHONE RECEIVERS. The high frequency at which the fundamental mechanical resonant frequency of a small Rochelle salt crystal occurs permits its use as a motor or motor and diaphragm in a telephone receiver. The large variation in normal impedance with frequency makes the response more dependent on the source impedance than it is when the normal impedance is constant.

11. PERFORMANCE AND TESTS

One criterion of the excellence of a complete sound-transmission system is its ability to produce the same space-pressure-time pattern at the ears of the listener that would be experienced if he were immersed in the sound field in the region of the original sound. This can be achieved with substantially complete realism in a binaural system in which two transmission links couple two properly mounted pressure microphones to two broadband telephone receivers placed on the ears of the listener.

In practice there is a growing tendency to apply a second criterion, namely the ability of a system to provide reproduction which is judged to be pleasant by a jury selected on the basis of scientific sampling of the public or the ultimate listening group. This second criterion has grown in importance for a number of reasons. One is that most listeners prefer loudspeakers to headphones, thus introducing complex effects due to coupling the loudspeaker to the ears of the listener via the listening room. Probably the most important reason for choosing pleasantness as the criterion of excellence is that radio and phonograph systems have in large measure attained the stature of an art medium in their own right. This has led to extensive work on the design of new types of studios and the development of new pickup techniques, all intended to please the listener in his home.

Though the complete test of a speaker is a complicated process, all the objective measurements have as their goal a satisfactory correlation between their results and those of properly conducted listening tests. Thus the ear becomes the final arbiter of the excellence of a sound system.

For discussing the aural performance it is desirable to use terms descriptive of the frequency ranges and of certain characteristics of the reproduced sound. As a group these terms constitute the imagery by which sound systems are described. They relate principally to two important frequency regions: 400 to 800 cps, which contains most of the energy of speech and music; and 2000 to 3000 cps, where the energy components contribute most strongly to the loudness. The first region is located in what most listeners classify as the low middles. A speaker with normal low middles is said to have "body," while an excess is "muddy," and a deficiency sounds "thin." When the second range, in the lower highs, is normal, reproduced sound is "crisp"; an excess introduces "bite" and is "brilliant"; a deficiency reduces articulation and loudness and, if no extreme highs are present to provide "spit," may yield a "mellow" speaker.

The whole range is conveniently divided into the very low frequencies, below 100 cps, lows from 100 to 250 cps, lower middle from 250 to 800 cps, upper middles from 800 to 2000 cps, lower highs from 2000 to 4400 cps, middle highs from 4400 to 8000 cps, and extreme highs above 8000 cps. The boundaries are quite arbitrary and will depend on the auditor, listening environment, and program material.

OBJECTIVE PERFORMANCE CRITERIA. Loudspeaker performance is determined to a large extent by the response, efficiency, directivity, distortion, power capacity, and impedance. The relative importance of these and other less important performance criteria depends on the application involved.

The response is that characteristic of the acoustic output, expressed as a function of frequency, to which a particular aural effect is closely proportional. The type of response will depend on the use to which the speaker is put. For speakers used outdoors or for very directional speakers the free space axial sound pressure level, as a function of frequency, is used and is called the sound pressure response.

For use indoors the sound pressure level is determined by the total acoustic power radiated and by the shape, size, and acoustical impedance of the room and its contents. Thus the effect of the speaker alone is approximated by the total radiation, usually ascertained by integration of the sound pressure measured in free space at fixed radius vector and variable angle. The total radiation response is modified by the "room response" to produce the desired effect on the ear, and it is often helpful to measure the rms average sound pressure in a room which is representative for the intended applications of the speaker. This mean pressure (corresponding to the mean energy density) response has good correlation with the aural effect when the measuring room is also that used for listening.

In Fig. 8 are shown the sound pressure, total radiation, and mean energy density response curves for a direct radiator speaker in a finite baffle. The excess of high frequencies

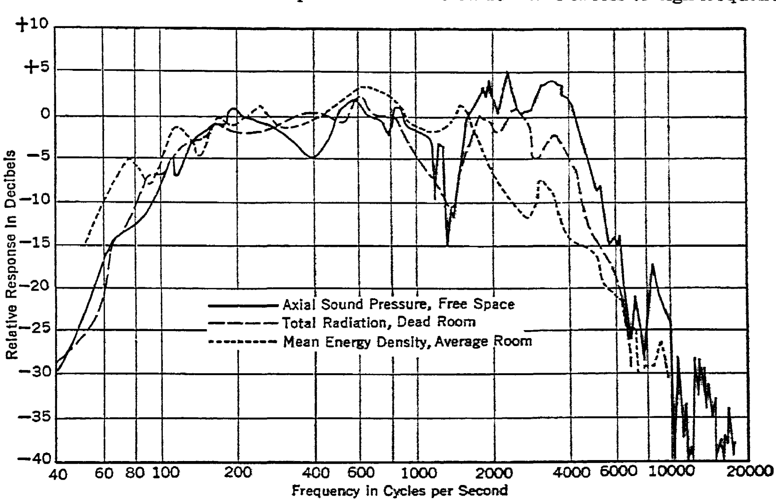


FIG. 8. Three Types of Response-frequency Curves for a Direct Radiator Speaker in a Finite Baffle

in the pressure response is due to directionality, and the difference between the second and third is due to the "response" of the room.

The system efficiency of speakers has been discussed in article 5. The loudness efficiency of two speakers is compared using a generator of constant emf and impedance, attenuation being inserted in the amplifier for the louder speaker. This judgment is usually the first of any listening test. It is becoming standard practice to test speakers with a source whose internal impedance is equal to its rated load impedance. The basis of stating the input to the speaker is conveniently the maximum power that may be taken from such a source, and it is called the power available to the speaker.

The series of sound pressure-frequency response curves run at various azimuths depict the directional characteristics of the speaker. Another clue to the directivity is obtained by comparing the sound pressure and total radiation response curves (see Fig. 8). When the wavelength is long compared to the size of the radiator, the radiation is uniform, and the curves should coincide. The differences at higher frequencies then indicate the magnitude of the directional effects.

If the listener's ear is relied on completely to select speakers having high loudness and (apparent) bass and treble response, there usually result speakers having large amounts of non-linear, transient, and frequency response distortion. The need for quantitative tests is indicated by Fig. 9, which shows the non-linear distortion in the sound pressure of a typical 12-in. speaker mounted in an 8 by 8 ft baffle. The speaker has an equivalent piston diameter of 10 in., a high force factor, and a voice coil $^7/_{16}$ in. long moving in a gap with a $^3/_8$-in. top plate.

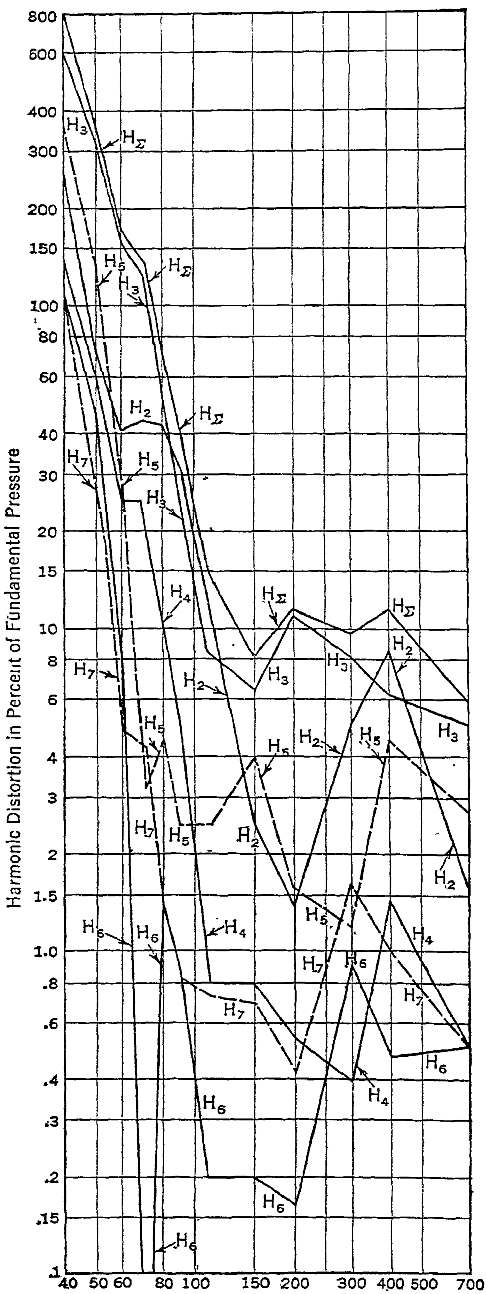


FIG. 9. Steady-state Non-linear Distortion in Moving-coil Direct Radiator Speaker

The curves are typical of commercial direct radiator speakers. Approximately half the distortion at low frequencies is due to the non-linear force displacement curves of the diaphragms. The remainder is due to the non-linear average-flux-displacement curves of the driving mechanism.

Since these curves are representative of commercial practice they are of interest in indicating the amount of distortion the untrained listener will tolerate. Speakers designed to give less distortion for a given cost will have less loudness efficiency, and less apparent bass response.

Distortion curves of horns show distortion of the same order in the middle frequency range. About a third of an octave above the theoretical cutoff frequency, the distortion begins to rise, and below cutoff usually considerably exceeds the distortion for direct radiator speakers. This distortion largely accounts for the horn's sounding as if it radiated effectively below its cutoff frequency.

For a more complete discussion of testing methods see the references in the bibliography marked "tests."

Another measure of distortion is the intermodulation of two signals of non-harmonic frequencies producing inharmonic modulation products. Because the annoyance of this type of distortion is extremely high, much smaller amounts may be tolerated than with non-linear distortion. Though the technique is not yet standardized, a useful procedure is to fix a low-frequency signal at a large value while a variable high-frequency scanning signal of much smaller value is applied through a linear combining circuit. A plot of the sound pressure due to the scanning signal and that due to the modulation products affords a means of assessing the audibility of the modulation products. It is important that this be done as a continuous function of frequency in order not to miss regions of large distortion. This is shown in Fig. 10 for an experimental direct radiator speaker with a 12-db difference between power available for low- and high-frequency signals.

The power rating of a loudspeaker is intended to provide the user with a yardstick by which he may set operating conditions to insure trouble-free life at any input up to the maximum. While this rating may be limited by distortion, thermal overload, or mechanical overload, the latter two are usually controlling. Thus the power rating usually sets a limit above which mechanical failure may be expected to occur. In one method of specifying this power, a saturating signal of a variety of program material is supplied to an amplifier of adjustable saturation power, used as the source. The rating power is that below which failure will not occur within, say, 100 hours for at least 90 per cent of a group of speakers. Thus the rating is essentially in terms of an amplifier whose output will be safely handled. In another method of rating, the amplifier has a much greater rating power than the speaker and is supplied a synthetic signal, such as noise, warble tone, or multitone. These have reasonably constant spectral composition and peak-to-rms ratio, and the power is increased until the 100-hour failure point is reached. Though neither method is yet standardized, it is nevertheless the intent of both to provide a realistic and useful value.

A final important characteristic of a speaker is its impedance. Both modulus and angle usually exhibit considerable variation with frequency, and the problem is to connect this load to a resistive source of essentially constant impedance. When properly connected, the source is able to deliver to the load the maximum program energy consistent with distortion requirements. In practice this "connection" impedance is selected by the manufacturer on the basis of his listening tests and experience. It may be called the rating impedance of the speaker, and it indicates the tap on the amplifier output to which the speaker should be connected. It is not necessarily the value at 400 cps, or the minimum in direct radiator speakers. Besides its use in determining speaker connections, the rating impedance is often used as the generator resistance in speaker tests.

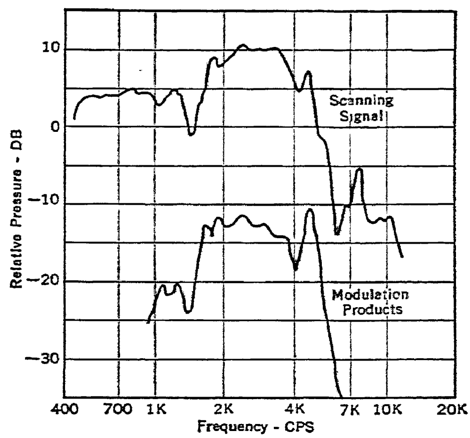


FIG. 10. Modulation Products for Experimental 12-in. Direct-radiator Speaker in Infinite Baffle. Available power input, 100 cps fixed, 16 watts; scanning signal, 1 watt.

BIBLIOGRAPHY

American Standards Association, *Standard* C16.4-1942. (Tests.)

Ballantine, S., *Proc. I.R.E.*, Vol. 23, 618 (1935). (Tests.)

Barnes, E. J., *W.E.*, Vol. 7, 248, 301 (1930). (Tests.)

Barrow, W. L., *J.A.S.A.*, Vol. 3, 562 (1932).

Bostwick, L. G., *J.A.S.A.*, Vol. 2, 242 (1930). (High-frequency horn unit.) *B.S.T.J.*, Vol. 8, 135 (1929). (Speaker tests.)

Von Braunmühl, H. J., and W. Weber, *Akus. Zeits.*, Vol. 2, 135 (5/37). (Annoyance of distortion.)

Chinn, H. A., and P. Eisenberg, *Proc. I.R.E.*, Vol. 33, 571 (9/45). (Listener preferences.)

Cosers, C. R., *W.E.*, Vol. 6, 353 (1929). (Hornless speaker.)

Davis, A. H., *Modern Acoustics*. (General.)

Edelman, P. E., Condenser Loud Speakers with Flexible Electrodes, *Proc. I.R.E.*, Vol. 19, 256 (1931).

Fletcher, H., *Proc. I.R.E.*, Vol. 30, 266 (6/40). (Role of hearing.)

Gerlach, E., *Physik. Zeit.*, Vol. 25, 675 (1924). (Ribbon speaker.)

Greaves, V. F., F. W. Kranz, and W. D. Krozier, The Kyle Condenser Loud Speaker, *Proc. I.R.E.*, Vol. 17, 1142 (1929).

Hahnle, W., *Wissenschaftliche Veröffentlichungen aus dem Siemens Konzern*, Vol. 10, 73 (1931), also Vol. 11, 1 (1932). (Hornless and condenser speaker.)

Hanna, C. R., *Proc. I.R.E.*, Vol. 13, 437 (1925). (Balanced-armature magnetic speaker.) *J. A.I.E.E.*, Vol. 47, 607 (1928). (Horn unit.) Theory of the Electrostatic Loud Speaker, *J.A.S.A.*, Vol. 2, 143 (1930).

Hartmann, C. A., and H. Jacoby, *Elek. Nach. Tech.*, Vol. 12, 163 (1935). (Tests.)

I.R.E. Standards on Electroacoustics (1938).

Kellogg, E. W., *J.A.S.A.*, Vol. 2, 157 (1930) (speaker tests); Vol. 3, 94 (1931) (low-frequency radiation); Vol. 4, 56 (1932).

Kennelly, A. E., *Electrical Vibration Instruments*. (Tests.)

Knowles, H. S., *Electronics*, Vol. 4, 154 (1932); Vol. 6, 240 (1933). (Tests.)

McLachlan, N. W., Speakers, *J.A.S.A.*, Vol. 5, 167 (1933) (tests, speakers, and bibliography); *W. E.*, Vol. 9, 329 (1932), Vol. 10, 204, 375 (1933) (distortion.)
Meyer, E., *Electroacoustics*. (Tests, general.)
Monch, W., *Mikrophon und Telephon*.
Neumann, H., *Zeit. tech. Physik*, Vol. 11, 548 (1929); *Wissenschaftliche Veröffentlichungen aus dem Siemens-Konzern*, Vol. 9, 226 (1930); and *Siemens Zeitschrift*, Vol. 10, 562 (1930). (Speakers and tests.)
Oliver, D. A., *W. E.*, Vol. 7, 653 (1930); Vol. 10, 420 (1933).
Olney, B., *Proc. I.R.E.*, Vol. 19, 1113 (1931) (speaker tests); *J.A.S.A.*, Vol. 8, 104 (10/36) (enclosures).
Olson, H. F., *Acoustical Engineering*. (Construction, tests, general.)
Olson, H. F., and F. Massa, *Applied Acoustics* (speakers-tests); *J.A.S.A.*, Vol. 6, 250 (1935) (telephone receivers); *Proc. I.R.E.*, Vol. 21, 673 (1933) (telephone receivers.)
Pedersen, P. O., *J.A.S.A.*, Vol. 6, 227 (1935); Vol. 7, 64 (1935). (Subharmonics theory.)
Rice, C. W., and E. W. Kellogg, *J. A.I.E.E.*, Vol. 44, 982, 1015 (1925). (Hornless m.c. speaker.)
Schafer, O., *Hochf. u. Elek.*, Vol. 44, 101 (1934). (Speaker test.)
Schaffstein, G., *Hochf. u. Elek.*, Vol. 45, 204 (1935). (Speaker and tests.)
Schottky, W., *Physik Zeit.*, Vol. 25, 672 (1924). (Ribbon speaker.)
Seabert, J. D., *Proc. I.R.E.*, Vol. 22, 738 (1924). (Speakers.)
Stenzel, H., *Handbuch der experimental Physik*, Vol. 17/2, 254 (1933) (loudspeaker, general); *Electrische Nach. Tech.*, Vol. 4, 239 (1927); Vol. 6, 165 (1929); Vol. 7, 90 (1930); also, *Zeit. tech. Physik*, Vol. 10, 569 (1929), and *Ann. der Physik*, Vol. 11, 947 (1930) (directivity of radiators).
Vogt, H., Elec. tech. Zeit., Vol. 52, 1402 (1931). (Test condenser speaker.)
Von Schmoller, F., *Telefunken Zeit.*, Vol. 15, 47 (1934). (Subharmonics.)
Wagner, K. W., *Die wissenschaftlichen Grundlagen des Rundfunkempfangs*. (General and speakers.)
Wegel, R. L., *J. A.I.E.E.*, Vol. 40, 791 (1921). (Moving-coil and magnetic-armature theory.)
Wente, E. C., and A. L. Thuras, *B.S.T.J.*, Vol. 7, 140 (1928) (horn unit); Vol. 13, 259 (1934).
Wheeler, H. A., and V. E. Whitman, *Proc. I.R.E.*, Vol. 23, 610 (1935). (Test.)
Williams, S. T., *J. Franklin Inst.*, Vol. 202, 413 (1926). (Horn speaker.)
Wolff, I., and L. Malter, *J.A.S.A.*, Vol. 2, 201 (1930). (Directivity of radiators.)
Wood, A. B., *Sound*. (General.)

MICROPHONES

By Hugh S. Knowles

A microphone is an electroacoustic transducer actuated by energy from an acoustical system and delivering energy to an electrical system, the spectral composition of the energy in the two systems being substantially equivalent. They may be classified as to reversibility, type of generator, acoustical quantity to which the output is proportional, and type of directional characteristic.

Reversible microphones are those that are also capable of converting electrical into acoustical energy, and they include moving-conductor (ribbon and moving-coil), variable-capacitance (condenser), variable-reluctance (moving-iron) and piezoelectric (crystal) generator types. The most important irreversible microphone is the carbon (variable-resistance) type.

Reversible microphones may be treated analytically by the same equations used for the corresponding loudspeakers, provided that the emf in the electrical mesh is set equal to zero and a force is introduced into the mechanical mesh (see Loudspeakers, p. 13-11).

12. FORCE ON THE MICROPHONE

In practically all microphones the electrical output is proportional to the net force on the generating element; less common types are those directly sensitive to the particle velocity (hot-wire) and to the temperature (pyroelectric) of the sound wave.

PRESSURE MICROPHONES. In pressure-responsive microphones the active surface is moved by a force which is the integral of the pressure over it. The pressure will deviate from that in the free wave owing to the effects of diffraction, cavity resonance, and mechanical impedance of the active surface. The last effect is usually small, and the others may be minimized or in some instances utilized by careful attention to the size and shape of the microphone. The microphone response may be stated in terms of the undisturbed pressure in the wave, or in terms of the pressure at the active surface; these are respectively the field and the pressure response.

Diffraction effects are important when the maximum dimension of the microphone becomes much greater than a tenth wavelength. For frequencies well above this region the pressure will be doubled (6 db) over the undisturbed value unless the diaphragm impedance is very low. Since diffraction effects are also functions of angle, large pressure-sensitive microphones may be expected to be directional at the higher frequencies. Screens are sometimes used to modify the diffraction effects.

When the active surface is recessed, the shallow cavity will exhibit a broad resonance which may be used to improve the high-frequency response. Though "gains" of the order of 5 db may be attained, the sensitivity to angle again introduces directivity. If the force resulting from a wave of normal incidence is to vary less than 2 db from that at grazing incidence, then the active surface should be recessed no more than 0.1 its diameter. At the highest frequency to be reproduced the diameter should be less than a half wavelength. In Fig. 1 is shown a miniature condenser microphone which approximates these conditions up to over 3 kc.

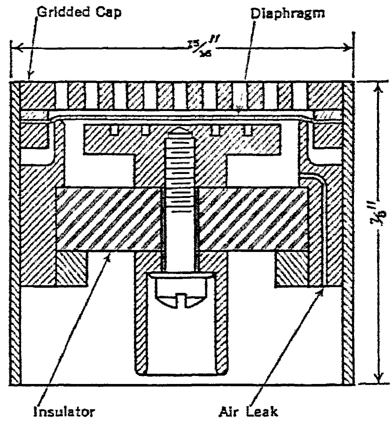


FIG. 1. Miniature Condenser Microphone

PRESSURE-DIFFERENCE MICROPHONES. When two points in space are separated by a small fraction of a wavelength the difference in pressure at these points approximates the product of the distance they are separated and the pressure gradient in the sound field. For this reason pressure-difference microphones are sometimes loosely called pressure-gradient microphones. (See also Ribbon Microphones, in article 13.)

When both sides of the active surface are exposed to a sound wave, the net force on it depends on the difference between the forces on the two sides. Let the acoustic path length from one side to the other be l_b. Then when a wave of undisturbed sound pressure amplitude, p, is incident on the microphone at an angle θ to the axis of symmetry, the magnitude of the net force, f, is, assuming no gradient along the active surface, and also assuming that the pressure difference may be replaced by the gradient,

$$f = 2S_d p \sin \frac{\omega l_b \cos \theta}{2c} \tag{1}$$

If ωl_b is sufficiently small,

$$f = \frac{S_d p \omega l_b \cos \theta}{c} \tag{2}$$

Thus, up to frequencies for which the path difference l_b is an appreciable fraction of a wavelength, the net force is proportional to that path difference and to the frequency. When the wave front is spherical, the pressure gradient exceeds its value for a plane wave by the factor $[1 + (c/\omega r)^2]^{1/2}$, which is small for ωr large. When the wave front is that due to a piston, the axial gradient contains terms dependent on the ratio of piston diameter to microphone distance. However, when this ratio is less than 0.5, the wave front is sufficiently spherical to permit the spherical wave correction factor to be used.

FLUID IMPEDANCE. Studies of reciprocity laws indicate that the force of the wave is applied through a "generator impedance" equal to that into which the active surface works when radiating the same wave fronts as those applied. See article 2 for results for some simple shapes. In general this added impedance is predominantly masslike up to the high-frequency region.

13. MOVING-CONDUCTOR MICROPHONES

In this class electrical energy is generated by the motion of single or multiple conductors in a magnetic field, as exemplified by ribbon and moving-coil types.

RIBBON MICROPHONES. The more important constructional details of a ribbon microphone are shown in Fig. 2. The moving ribbon element is both diaphragm and conductor; it may be 0.312 in. by 2 in. by 0.0002 in. and is corrugated to minimize spurious modes of vibration. As a low-frequency approximation, the effect of the fluid impedance is to add a mass of the order of magnitude of that of the ribbon; this added mass decreases at higher frequencies. The mechanical system is resonated at a frequency below the range to be reproduced, resulting in a mass-controlled device.

In pressure-difference responsive ribbon microphones, both sides of the ribbon are exposed to the wave. Thus the net force at low frequencies is roughly proportional to the pressure gradient and hence to the velocity in a plane wave. For this reason the term velocity is sometimes applied to microphones of this type. However, it appears desirable to reserve the term velocity to describe devices such as the Rayleigh disk in which the actuating force arises more directly from the particle velocity. Up to fairly high frequencies both the force and impedance (mass controlled) are proportional to frequency, leading to a constant ribbon velocity and hence constant generated emf. Control of the response beyond this point is achieved by adjusting the size and disposition of the pole pieces and other elements separating the two faces of the ribbon.

If one side of a ribbon microphone is shielded from the wave by a damped enclosure, the device becomes pressure-responsive. With a sound pressure independent of frequency the output is independent of frequency if the enclosure adds sufficient resistance to make the effective impedance of the diaphragm substantially resistive. Such a pressure-responsive ribbon microphone is used principally with a pressure-difference responsive type in a directional combination.

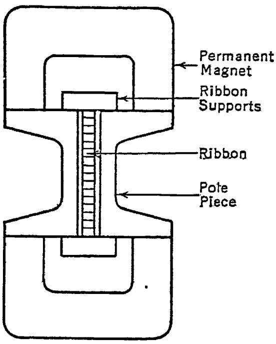


FIG. 2. Pressure-difference Ribbon Microphone.

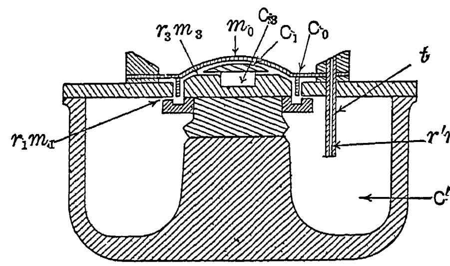


FIG. 3. Moving-coil Microphone

MOVING-COIL MICROPHONE. When the moving conductor is a moving coil, there results a form similar to that of a moving-coil loudspeaker, as in Fig. 3. With the net force in the mid and high frequency ranges largely due to that on the outside of the diaphragm, the device is pressure-responsive in these ranges. In order to obtain uniform coil velocity and hence response the moving system is designed to have substantially uniform impedance. The diaphragm and coil assembly is resonated in the low-middle frequency range.

Coupled circuits are used to maintain uniform response at the extreme frequencies. In Fig. 3 the circuit elements leading through the tube t to the rear of the diaphragm are so chosen as to achieve an increasing force at low frequencies to offset the increasing stiffness reactance of the diaphragm. At high frequencies the stiffness reactance of the air film under the dome and the impedance of the tuned cavity in the pole tip modify the impedance of the diaphragm and coil assembly so as to maintain approximately uniform velocity. Thus it is possible to attain a fairly uniform response from 40 to 10,000 cps.

It is possible to combine the output of adjacent pressure and pressure-difference microphones in such a manner that directional patterns are obtained which are members of the limaçon family of which the cardioid is a special case. Common types use moving-coil pressure and ribbon pressure-difference microphones connected through suitable phase and amplitude controlling networks.

14. CONDENSER MICROPHONES

Most variable-capacitance microphones are made pressure-responsive, with a mechanical system tuned above the desired range. A frequency-independent driving force acting upon this stiffness-controlled system results in a displacement independent of frequency. Since the output emf is closely proportional to the displacement, a flat pressure response is approached.

To the mass of the diaphragm is added the radiation mass, and to its stiffness, that of the air film between diaphragm and rear electrode. See Fig. 1. The diaphragm is of a high-tensile-strength aluminum or stainless-steel alloy of the order of 0.00075 in. thick, and is tuned by stretching, the smaller diaphragms being tunable to higher frequencies. The resistance of the thin interelectrode air film materially reduces the magnitude of the high-frequency resonant peak.

The small cavity and the high mechanical impedance of the condenser microphone make it particularly suitable for calibration by reciprocity methods. Two similar microphones whose comparative response is known are placed in a closed cylindrical volume of which each forms an end; the output of one is noted when the other is electrically driven.

As with all reciprocity methods, its success depends on the presence of a definite and calculable acoustic element, which here is the compliance of the common volume, corrected for microphone mechanical impedance. With careful control of acoustic and electrical parameters, excellent reproducibility of results is attained. The upper frequency limit may be extended to 15,000 or more cps by the use of hydrogen in which the wavelength is increased by a factor of nearly 4.

As reciprocity methods yield a pressure calibration, a correction for cavity resonance and diffraction is necessary. When the condenser microphone is used as a free field measurement standard, it is common practice to utilize these factors to extend the frequency range, with electrical equalizers added for smooth overall response.

15. MAGNETIC-ARMATURE MICROPHONES

A magnetic-armature microphone is a moving-iron (variable-reluctance) device which finds greatest application where ruggedness and high output are the main desiderata. An important use is in sound-power telephones (using no auxiliary source of power, such as batteries), which place a premium on maximum transmission of intelligence. This is achieved by resonating the moving system at two or more frequencies in the 1–3 kc band in which maximum articulation is obtained for a given amount of energy. (For force factor and force equations see article 7, Magnetic-armature Speakers.)

16. CRYSTAL MICROPHONES

In this type use is made of the piezoelectric properties of Rochelle salt, ammonium dihydrogen phosphate, or more rarely tourmaline and quartz. The sensitivity may be expressed in terms of the charge released for a given displacement of the driving point; this depends but slightly on temperature. However, the dielectric constant of Rochelle salt and hence the capacitance and emf generated changes considerably with temperature. The effect of this capacitance change is minimized by operating the crystal in a high-impedance circuit, or less frequently by shunting it with a fixed capacitance. The maximum operating temperature range is 40 to +130 deg fahr, and maximum output occurs at about 72 deg fahr.

The maximum operating temperature range of ammonium dihydrogen phosphate is −40 to +185 deg fahr. While the emf generated is greater than for Rochelle salt, the smaller dielectric constant results in such small capacitances that the full output is difficult to realize. The high resistances necessitated by the small capacitances also give rise to electrical noise problems.

Since the emf developed by crystals depends on the displacement, a stiffness-controlled system in which the displacement is substantially independent of frequency is used. When the microphone consists of a pair of composite plates arranged as in Fig. 4 to form a sound cell, the resonant frequency may range from 8 to over 40 kc.

When a single composite plate is clamped at three corners and the fourth is driven by a small diaphragm, the output voltage increases some 15 db over the sound cell construction, but the resonant frequency is then reduced to a few kilocycles. The main use of this type is for speech, in which case the output is allowed to rise substantially at the fundamental resonant frequency of 3 or 4 kc. Auxiliary acoustic meshes are used to obtain smooth response with reduced sensitivity in broad-band types used for music.

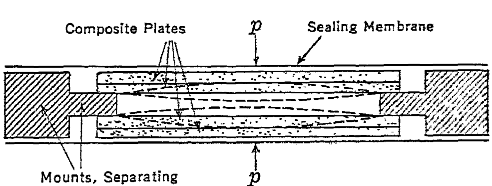


FIG. 4. Crystal Microphone ("Sound Cell" Type)

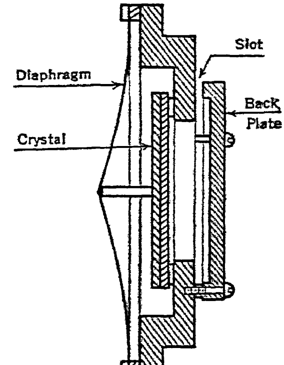


FIG. 5. Unidirectional Diaphragm-driven Crystal Microphone

A single transducer element crystal unidirectional microphone is shown in section in Fig. 5. When the slot is closed the microphone operates as a pressure type. (See Directional Characteristics, below.)

17. CARBON MICROPHONES

The carbon microphone, the most important irreversible microphone, is a variable-resistance type involving loose carbon contacts. The sound wave actuates a diaphragm which exerts a varying pressure on a large number of fine carbon granules, thus modulating a bias current obtained from a d-c source. Usually the resistance and output voltage are approximately proportional to the displacement, and so a stiffness-controlled system is employed. Because of the difficulty of obtaining uniform response, the present uses of carbon microphones are restricted to applications in which high output and "crisp" speech quality are paramount, as in military equipment. Another factor militating against its use for high-fidelity applications is the low signal-to-noise ratio, due to the random variation in contact resistance always present.

The small allowable size of the carbon button has been used to advantage in a close talking pressure-difference microphone with marked noise-reducing properties. Designed to be worn directly in front of the mouth, the microphone is immersed in the large pressure gradient field of the talker, while the ambient noise has a much smaller gradient. Thus the signal-to-ambient-noise ratio is improved, permitting intelligible communications from such noisy locations as planes, tanks, and engine rooms.

18. DIRECTIONAL CHARACTERISTICS

It is often important to collect sound arriving from a desired region to the exclusion of randomly incident sound, as in sound-reinforcing systems. The ability of the microphone to accomplish this is determined by the dependence on angle and frequency of the field response. There are many measures of directionality, such as the ratio of output for sound from the desired direction to the output for sound of random incidence of the same total power, the ratio of the output for sound of random incidence in the front hemisphere to that of random incidence in the rear hemisphere, or the ratio of the outputs at the angles of maximum and minimum response.

Most commonly used microphones may be classified as non-directional, bidirectional, and unidirectional. Since pressure is a scalar quantity, an ideal pressure microphone is non-directional. In practice complete freedom from directional effects is achieved only when the maximum transverse dimension is of the order of an eighth wavelength or less. Since most microphones are operated at frequencies up to 4000 or more cps, this requirement necessitates a microphone a centimeter or less in diameter. To provide adequate sensitivity microphones are made larger than the non-directional requirements dictate and commonly have a diameter of 2 to 5 cm. In pressure-actuated microphones this results in a non-directional microphone at low frequencies and a unidirectional one at high frequencies. Screens are sometimes mounted near the diaphragm to alter the sound field and make the microphone less directional.

The most common bidirectional microphone is the cosine pressure-difference or "gradient" type exemplified by the ribbon microphone discussed above.

Most unidirectional microphones have a directional characteristic, which, if taken in a plane through the principal axis, is given by $e_u = e_n + e_b \cos \theta$. If the microphone has separate non-directional (pressure) and bidirectional transducer elements, e_n and e_b are their respective voltages and e_u is their combined voltage. One type of unidirectional microphone is a pressure-difference type in which a single transducing element is used with a network to shift the phase of one of the pressures to alter the directional characteristic. In this type e_n corresponds to the voltage with no force contribution from the rear network and e_b to the voltage when the rear and front impedances are equal and the microphone is bidirectional. By adjusting the impedance of the rear network the relative values of e_n and e_b may be altered to give various directional characteristics. Single transducer microphones of this type are commonly made with crystal or moving-coil generators.

19. PERFORMANCE AND TESTS

The most important performance criteria are the field response (see Fig. 6), directivity, impedance, and inherent noise. Less important for general applications are the "dynamic" range (the range from minimum pressure, limited by electrical noise, to maximum pressure limited by non-linearity or by structural strength), the ratio of the responses for near and distance sources, termed the proximity index, and non-linear distortion.

OBJECTIVE TESTS. The field response of a microphone is a measure of the electrical output, for a specified frequency, when immersed in a plane progressive wave. When the effect of the impedance is considered, an expression of the form $20 \log E/p - 10 \log R$ results, in which E is the open-circuit emf, in volts, generated by the microphone; p is the undisturbed sound pressure, in dynes/cm^2; and R is the stated impedance of the microphone. This relation takes into account the effect of impedance in permitting the use of step-up transformers. For crystal and condenser microphones, R is sometimes assigned a value corresponding to the maximum stated value of transformer secondary impedance used with low-impedance microphones, usually from 30,000 to 100,000 ohms. This value permits a fairer comparison of all types in terms of the amount of amplification necessary, referred to an input grid.

The directivity is usually determined from field-response curves taken at various angles of incidence, or from the field response as a function of angle with the spectral composition of the test signal held constant. Single-frequency, narrow-band warbled or frequency-modulated and noise test signals are used. Unidirectional microphones should show greater than 10-db front-to-rear discrimination (15 db is a common value), while in the bidirectional (pressure-difference) type the front-to-side discrimination usually exceeds 20 db over a large frequency range.

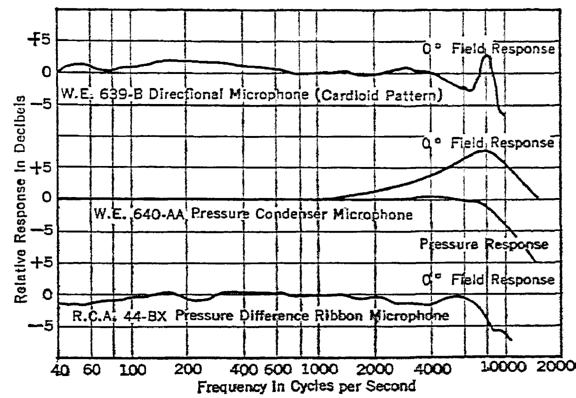


FIG. 6. Axial Response-frequency Curves of Three Microphones

To minimize frequency discrimination and reduce electrostatic and electromagnetic induction in long microphone lines, broadcast-type microphones have impedances ranging from 35 to 250 ohms, with a trend toward 150 ohms. These microphones are usually operated into impedances ten or more times that of the microphone's. This results in about a 3-db improvement in the signal-to-inherent-electrical-noise ratio over that obtained when the load resistance equals the modulus of the microphone impedance.

The present technique of primary calibration is by means of the reciprocity technique mentioned under Condenser Microphones, article 14. This yields a pressure calibration, from which the field response may be derived by calculation or measurement (with scaled models) of diffraction and cavity resonance effects. A microphone thus calibrated may be used as a standard from which the response of other microphones may be obtained by comparison, in absolute terms. The source may be a sufficiently distant loudspeaker with very smooth response; or for special purposes an artificial voice is used to obtain field shape and diffraction effects approximating those of a person speaking. Although it is possible to obtain free-field reciprocity calibrations, much work remains to be done before the precision and stability of the pressure reciprocity method can be attained.

SUBJECTIVE (AURAL) TESTS. As with loudspeakers, the final acceptability of a microphone depends on subjective tests. A live or artificial voice may be used as a source for two microphones being compared, the outputs being alternately connected to an audition system. Such qualities as naturalness, smoothness, presence, articulation, the objectionability of intermodulation distortion, loudness efficiency, and transient distortion may best be compared by aural tests with a trained jury.

BIBLIOGRAPHY

Am. Stds. Assoc. Standard Z24.4-1938. (Calibration.)
Am. Stds. Assoc. Standard Z24.1-1942. (Terminology.)
Baerwald, H., *J.A.S.A.*, Vol. 12, 131 (1940). (Noise.)
Ballantine, S., *J.A.S.A.*, Vol. 3, 319 (1932). (Calibration.)
Bauer, B. B., *J.A.S.A.*, Vol. 13, 41 (1941). (Unidirectional microphone.)

Braunmühl, H. J. von, and W. Weber, *Hochfrequenztechnik u. Elektroakustik*, Vol. 46, 187 (1935). (Directional pressure-difference condenser microphone.)
Cook, R. K., *J. Research, National Bur. Stds.*, Vol. 25, 489 (1940). (Reciprocity.)
Crandall, I. B., *Phys. Rev.*, Vol. 11, 449 (1918). (Condenser microphone.)
Ellithorn, H. E., and A. M. Wiggins, *Proc. I.R.E.*, Vol. 34, 84 (1946). (Gradient microphones.)
Foldy, L. L., and H. Primakoff, *J.A.S.A.*, Vol. 17, 109 (1945). (Reciprocity.)
Gerlach, E., and W. Schottky, *Physik Zeits.*, Vol. 25, 276 (1923). (Ribbon microphone.)
Glover, R. P., *J.A.S.A.*, Vol. 11, 296 (1940). (Cardioid unidirectional microphones.)
Harrison, H. C., and P. B. Flanders, *B.S.T.J.*, Vol. 11, 451 (1932). (Miniature condenser microphone.)
Marshall, R. N., and W. R. Harry, *J. S.M.P.E.*, 33, 54 (1939). (Cardioid microphone.)
Massa, F., *J.A.S.A.*, Vol. 17, 29 (1945). (ADP microphone.)
Olson, H. F., *Elements of Acoustical Engineering*. (General.)
Olson, H. F., *R.C.A. Rev.*, Vol. 6, 36 (1941). (Reciprocity.)
Sawyer, C. B., *Proc. I.R.E.*, Vol. 19, 2020 (1931). (Crystal microphones.)
Schottky, W., *Zeits. f. Physik*, Vol. 36, 689 (1926). (Reciprocity.)
Wente, E. C., and A. L. Thuras, *J.A.S.A.*, Vol. 3, 44 (1931). (Moving-coil microphones.)

MAGNETIC RECORDING AND REPRODUCING OF SOUND

By L. Vieth and H. A. Henning

The earliest record of magnetic recording is credited to Poulsen, who in 1900 described his "Telegraphone." Since that time, though much has been learned and vastly improved results have been obtained, extensive use of the method as a recording process of great potentialities was not achieved until the war years of 1941–1945.

The magnetic recording method possesses certain unique advantages. The mechanism may be simple to operate and rugged. The record can be played immediately after recording and can be replayed practically any number of times. The recording may be erased and the medium reused as often as desired.

Magnetic recording involves three fundamental operations or processes: (1) erasing; (2) recording; (3) reproduction or playback. Erasing is the process by which the magnetic recording medium is either neutralized or saturated to obliterate any signal previously recorded. In the early phases of development saturation by a strong d-c field was used. More recently neutralization of the medium by a high-frequency a-c field has been required in conjunction with modern methods of recording with an a-c bias. Recording is accomplished by applying, through the recording head, the signal to be recorded superimposed on a biasing current. The bias current may be either direct or high-frequency alternating current. Considerations governing the choice of biasing and erasing methods are discussed in article 21. The bias current is necessary to obtain a faithful recording of the impressed signals. The medium, as it passes under the influence of the recording head, is magnetized in proportion to the variations of the signal current, and this magnetization remains until erased. Reproduction or playback is accomplished by passing the recorded medium over a magnetically sensitive head, usually at the same velocity as in recording. The voltages induced in the head are then amplified and equalized to obtain the desired frequency characteristic. All these processes may be accomplished with the same head at some sacrifice in performance.

20. ERASING, RECORDING, AND REPRODUCING ARRANGEMENTS

Physical arrangements by which the magnetic forces are applied to a magnetic medium (usually in the form of a tape or wire) have taken several forms. They may be divided into three broad classifications: (1) perpendicular magnetization, in which the direction of magnetization is normal to the surface of the medium; (2) modifications of (1), which alter perpendicularity somewhat; and (3) longitudinal magnetization in which the direction of magnetization is parallel with the direction of motion of the medium. Transverse magnetization in which the direction of magnetization is parallel to the surface of the medium and normal to the direction of motion is a fourth possible classification which will not be considered here. Figure 1 illustrates schematically the three arrangements. Each consists essentially of cores of high-permeability material surrounded by one or more coils. Figure 1(*a*) shows the perpendicular application of magnetic force to a medium in the form of tape, moving between the poles, each consisting of one lamination in exact alignment. Each pole is surrounded by a coil. A concentration of recording flux is obtained by the small thickness of lamination in the direction of tape travel.

Figure 1(*b*) illustrates a more efficient and practical modification of (*a*). Much thicker cores are used, which are in contact with opposite surfaces of the tape. Flux concentration is accomplished by offsetting the poles in the direction of travel by an amount almost equal to their thickness. A coil surrounds the advance pole. The tape is therefore magnetized in a preponderantly perpendicular manner with a longitudinal component. Induction at short wavelengths is most pronounced on the side of the tape in contact with the coil-bearing pole. Omission of the coil on the receding or following pole avoids a secondary concentration of recording flux as the tape leaves the influence of the recording head, thus avoiding modulation of the already recorded signal. The schemes illustrated in 1(*a*) and 1(*b*) require that intimate contact must be maintained by pressure against both surfaces of the tape. Joints must therefore be carefully made, and the two surfaces must be uniformly parallel to insure good contact at all times. Furthermore, these two schemes are of limited use on wire and are even less useful on non-magnetic tapes coated on one side with magnetic materials. Figure 1(*c*) shows an arrangement which applies magnetizing force longitudinally. Contact is made on only one side, and the principle is therefore conveniently applicable to all forms of recording media. A ring of permeable material

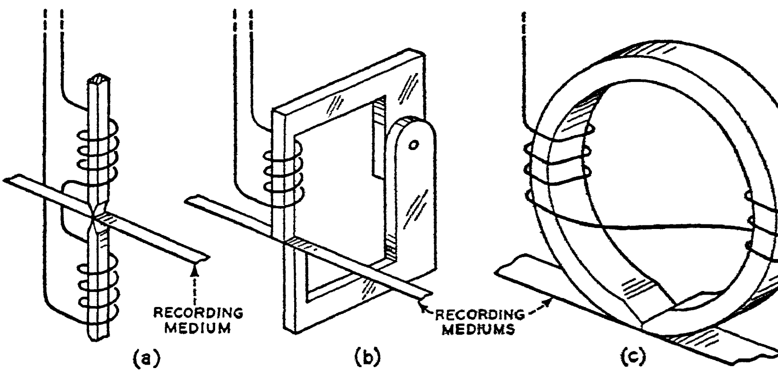


FIG. 1. Typical Magnetic Recording and Reproducing Arrangements

wound with a coil is provided with a small air gap at the point where it touches the recording medium. A portion of the flux in the air gap passes through the medium. The resulting magnetization is substantially longitudinal. The concentration of flux is accomplished by the use of very small gaps ranging from 0.003 in. to less than 0.0003 in., depending on the medium.

Although all the above arrangements may be used for erasing, recording, and reproducing, it is expedient to alter their characteristics in the interests of one process or another. Separate heads are frequently provided for each process. In each case it is important that the magnetic action of the head in recording and reproducing be concentrated along a line perpendicular to the direction of motion. The greatest concentration of active flux lines is obtained for a given pole-piece arrangement when the tip reluctance of the recording or reproducing pole piece is low. Poles or cores are usually laminated to reduce the frequency discriminating effects of eddy currents.

21. ERASING, RECORDING, AND REPRODUCING PROCESSES

RECORDING SPEED. The various systems used for driving the magnetic media will not be discussed here. There are, at present, no standardized recording speeds, and in most applications the medium is moved past the recording and reproducing heads at the slowest speed that will insure the desired high-frequency response. In practice these speeds range from a few inches to several feet per second. In order to obtain a frequency range of 10,000 cycles it is modern practice to employ recording speeds between 1 $^1/_2$ and 5 ft per sec, depending on the properties of the recording medium and the methods of recording.

ERASING. The process used to obliterate a previously recorded signal on a magnetic medium is called erasing. In most applications the erasing operation is performed during the recording process by a head located somewhat in advance of the recording head. There are, however, instances in which it is more convenient to perform the erasure as an independent operation.

A previous recording may be removed either by saturating the medium with a strong d-c field or by neutralizing the medium with an a-c field of diminishing intensity. In the first process every portion of the medium entering the erasing head is carried to saturation or a degree of magnetization exceeding the strongest recorded signal. The erasing head may consist of a permanent magnet or an electromagnet supplied by direct current. In certain designs the configuration of leakage flux lines around the head can cause a second and reversed field to act on the medium as it leaves the head. Under such conditions the initial saturation is followed by a partial demagnetizing operation, and the ultimate state of the medium may be considerably less than fully saturated. As the optimum d-c recording bias is determined by the magnetic state of the erased medium it is apparent that the design of the erasing head influences the constants of the recording process.

It is now common practice in high-quality magnetic recording to erase by a neutralizing or demagnetizing action. This is accomplished by subjecting each element of the medium to a cyclicly varying magnetic field whose maximum intensity between the poles of the head produces saturation. As the element moves away from the head it is subjected to a continuously diminishing cyclic magnetization which leaves the medium in a neutral state. A magnetizing force approximately three times the coercive force of the medium has been found to be satisfactory. A field intensity of 1500 oersteds may be required to demagnetize some of the modern high coercive force materials.

Some use has been made of an air-core coil for erasure of the lower-coercive-force materials. The medium is drawn through the center of the coil, and a comparatively low-frequency erasing current is required. If erasing is accomplished with the recording head a higher-frequency current is used to provide the necessary number of reversals in the short time the medium is within the influence of the head.

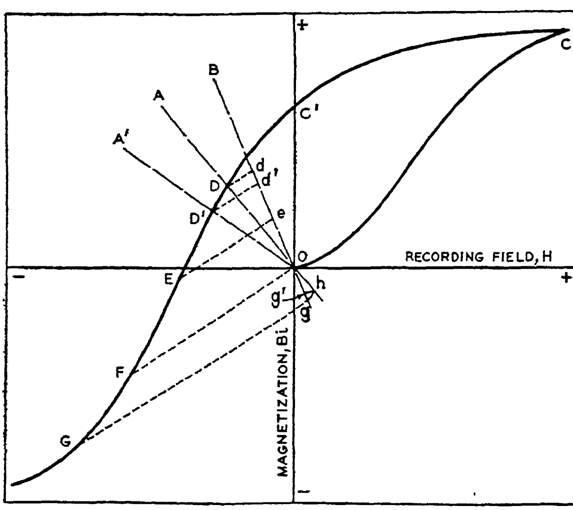


Fig. 2. Magnetic History of an Element of Recording Medium. Saturation erase, d-c Bias.

In certain instances a combination of d-c saturation followed by partial neutralization has been employed. In such circumstances the recorded signal is successfully erased but there remains a residual d-c component on the medium which is responsible for somewhat increased distortion.

RECORDING. The process of recording consists of impressing on the moving magnetic medium a varying induction which is directly proportional to the instantaneous value of the recording head current. The recording head is usually associated with an amplifier so designed that the recording currents are independent of the impedance characteristic of the head. In most applications recording currents are small and very little power is required. In certain applications electronic amplification is not used and the recording current is obtained directly from a carbon microphone.

The linear relationship between recording current and the resulting magnetization on the medium is obtained by the use of a superimposed biasing field, usually by superposing a biasing current on the recording current fed to the head. Without the bias the reproduced signals are weak and very distorted. The biasing field is adjusted to a magnitude that minimizes the distortion without introducing unnecessary background noise. The proper value is usually slightly less than that which will yield the strongest reproduced signal. Two biasing methods are available. It is now customary to use a high-frequency (supersonic) biasing current, and the greatest dynamic recording range may be obtained by this method. In earlier applications a d-c biasing current was used; it is advantageous where extreme compactness and simplicity of equipment are essential.

The magnetic processes involved during recording may be illustrated by hysteresis loops such as those of Figs. 2, 3, and 4. These curves represent *Bi* (or *B-H*) vs. *H*, or intrinsic induction vs. magnetizing force. They thus represent the fundamental characteristic of the recording medium. Each element of the medium, as it passes the recording point, is subjected to a magnetizing field which is proportional to the algebraic sum of the instantaneous recording current and the biasing current. The element is then withdrawn from the field. For long uniformly magnetized sections the resultant state of each element of the section would be represented by a point on the *Bi* axis. The magnetic state of an element of a short magnet (because of self-demagnetization) is represented by a point on an appropriate demagnetization coefficient line. The demagnetization coefficient line *OA* of Figs. 2, 3, and 4 may be called the open circuit line, and it defines the state of the element when removed from the head. Whenever the element under consideration is within the playback head, the self-demagnetizing field is partially removed and the magnetic state of an element moves along a reversible minor hysteresis curve which may be represented by a straight line such as *Dd* in Fig. 2 to a steeper demagnetization coefficient line *OB*. The slope of line *OB*, which may be called the closed circuit line, is determined primarily by the reluctance present at the contact point of medium and head and thus is substantially independent of the recorded wavelength. In longitudinal recording the slope of the open-circuit line *OA* is to some extent a function of the recorded wavelength, and thus the self-demagnetization is a function of frequency. The effect on the frequency response of this action may be at least partially offset[4] by the use of high-coercive-force materials.

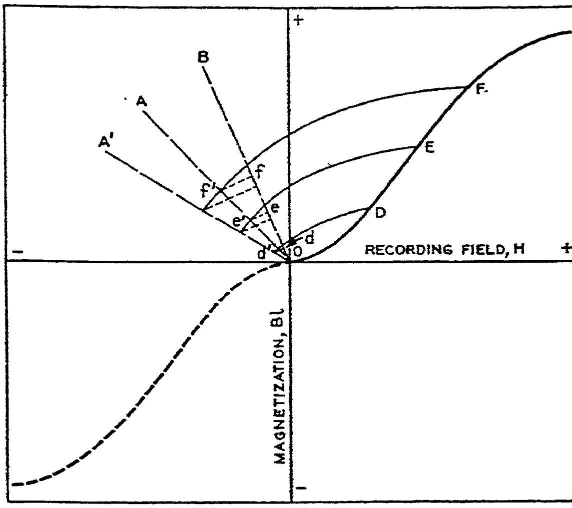


FIG. 3. Magnetic History of an Element of Recording Medium. Neutralization erase, d-c bias.

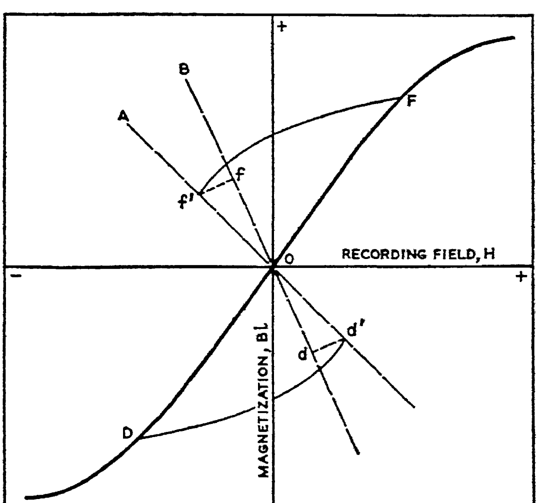


FIG. 4. Magnetic History of an Element of Recording Medium· Neutralization erase, a-c bias.

Figure 2 illustrates the magnetic process of recording with a d-c bias on a medium that has previously been saturation erased. During the erasing process the saturating field brings the magnetic state of each element of the medium to a point *C* on the hysteresis curve. Assuming that the element leaves the erasing field without being subjected to a reverse magnetization, the magnetic state of the element moves along the hysteresis curve *CC'D* to point *D* on the open-circuit line *OA*. Subsequently, when in the recording head, the state of the element is represented by point *d* on closed-circuit line *OB*, and as the minor hysteresis curve *Dd* traversed during this process is substantially reversible any applied recording field operating within this region leaves no change in impression on the medium. If after saturation a d-c bias is applied, the state of the element is brought to point E along the major loop, and this point is chosen to represent the midpoint of the available operating range. Variations in the recording field will then produce corresponding changes in magnetization. The combined action of the biasing field and the signal field then operates over a substantially straight section DEF of the hysteresis curve. When the element emerges from the recording head, its magnetic state returns along the minor loop to the open-circuit line OA and subsequently, when in the reproducing head, to the closed-circuit line OB along the same minor loop. Depending on the signal strength, the path of this action is along nearly straight, parallel, and reversible minor hysteresis curves lying between Dd and FO. Thus the flux induced during reproduction, for the state of magnetization corresponding to a point on the closed-circuit line OB lying between O and d, is practically linearly related to the applied recording field.

In Fig. 2, the point F represents one of the limits of the recording field which may be applied to the medium without serious distortion. This is because minor hysteresis curves have greater curvature after crossing the Bi axis and a minor hysteresis curve such as Gh leaves the magnetic state of the element at g' instead of the required point g.

The operating range during recording at short wavelengths is also restricted wherever the slope of the open-circuit line OA is a function of the wavelength. This is illustrated in Fig. 2 by line OA', which might be the slope of the open-circuit line in a medium longitudinally recorded at a short wavelength. The magnetic state of a saturated element is then represented by the point D'. The recording field is thus restricted to operation over the region between D' and F.

Figure 3 illustrates the magnetic process involved when a d-c bias is used with a completely demagnetized medium. The recording bias alone serves to carry the magnetic state of the element along the normal magnetization curve $ODEF$ to point E. The recording signal operates about E, which is chosen as the midpoint of the linear part of the magnetization curve. Upon leaving the recording head the magnetization of the element drops to line OA along one of substantially parallel hysteresis curves similar to and lying between Dd' and Ff'. The process is thereafter similar to that described for saturation erase. If, at shorter wavelengths, the slope of the open-circuit demagnetization coefficient line OA is reduced to OA', points d and f on the closed-circuit demagnetization coefficient line OB are proportionately reduced. This results in an attenuation of the recorded signal. It may be observed in Figs. 2 and 3 that the degree of magnetization of an element which has been biased but is otherwise unrecorded is substantially the same in either process, and also that the undistorted operating range is nearly identical. Because of these conditions the background noise and dynamic range in either process are substantially the same, and thus when operating with a d-c bias there is little advantage in providing extra equipment to demagnetize the medium.

It is now customary to erase by demagnetization and employ a high-frequency (supersonic) biasing current during recording. The action of the high-frequency bias has been variously explained. It is sufficient to say here that it serves the purpose of straightening out the bend in the normal magnetization curve around the origin. The resulting action is as shown in Fig. 4. The recording signal may then operate about the origin O and along the entire straight portion of the curve between D and F. The remainder of the process may be considered to be similar to that described for recording with a d-c bias on a demagnetized medium except that the recording range is doubled. This represents a doubling of the reproduced signal without an equivalent increase in background noise.

There is an even greater increase in signal-to-noise ratio inasmuch as the elimination of the d-c biasing component of magnetization serves to reduce the observable background noise. The biasing frequency employed in this method of recording is customarily quite high and usually leaves no permanent impression on the magnetic medium. Wherever a biasing frequency is recorded, it is beyond the range of the playback equipment and is not reproduced.

In common with all methods of sound recording, magnetic recording introduces a certain amount of distortion although modern magnetic recording systems are considerably improved in this respect. Distortion is introduced during the recording process from several sources. The most common source of distortion arises from the non-linear properties of the recording medium. It is apparent from Figs. 2 and 3 that there is an optimum value of d-c bias which will minimize curvature distortion. When using an a-c recording bias there is a minimum value which is sufficient to eliminate the curvature of the normal magnetization curve around the origin. Higher values than the necessary minimum may be used, but there is no advantage to be obtained. Distortion resulting from intermodulation between the bias frequency and the signal frequencies and their harmonics may be kept outside the operating range by using a sufficiently high biasing frequency. When recording with an a-c bias, distortion is increased by the presence of a residual d-c magnetization which may be present either in the recording and reproducing heads or in the erased medium.

A distortion may also be introduced by the recording head if the magnetizing field at the head is not sufficiently concentrated. At higher frequencies it is then possible for the recording signal to change while an element is within the influence of the head, and the element will retain an impression of the strongest magnetizing force to which it is subjected. The use of narrow recording gaps and modern high-permeability pole pieces has greatly reduced this type of distortion.

REPRODUCTION. In the reproducing process the varying magnetization impressed on the medium during recording produces a corresponding flux in the reproducing head. A portion of this flux threads the coil, and its variation due to the motion of the medium induces the signal voltage. Signal voltages are normally quite low, and considerable amplification is required.

The process by which the reproducing head picks up the signal flux is essentially similar for any of the heads of Fig. 1. The return path for all the signal flux in a magnetized element approaching the reproducing head is through the air. The magnetic state of such an element is described in Figs. 2, 3, and 4 by a point on an open-circuit demagnetization coefficient line such as OA. As the element approaches the head some of the flux which leaves the medium passes to the head. In the case of the head of Fig. 1(a) a portion of this flux threads the coil. The percentage of flux which threads the coil increases very rapidly as the element comes directly under the thin lamination and decreases as rapidly thereafter. Thus the total flux threading the coil is contributed by the magnetized elements under the thin lamination and those immediately adjacent. In the case of the offset pole pieces of Fig. 1(b) the major contribution to the flux threading the coil is made by the elements in the immediate region of the overlap point. Elements in contact with one pole tip and remote from the overlap point are either insufficiently coupled or, at short recorded wavelengths, are short-circuited by the pole tip and therefore do not contribute to the flux threading the coil. The ring-shape head of Fig. 1(c) operates in a somewhat similar manner. The head serves as a return path for flux leaving the surface of the longitudinally recorded medium. With the exception of flux from elements of the medium in the immediate vicinity of the air gap, the return path does not include the reproducing coil. Thus the flux which does thread the coil may be considered as being contributed by the elements in the immediate region of the air gap.

When the recorded element is in position at the reproducing point its magnetic state is described in Figs. 2, 3, and 4 by a point on the closed-circuit demagnetization coefficient line OB. The slope of this line is determined by the total reluctance of the return path for the element in question, and this reluctance is affected by the amount of contact between the medium and pole pieces. The flux threading the coil is then equal to the product of the Bi intercept of the point on OB, the normal cross-sectional area of the contributing elements, and an appropriate leakage factor. The contact reluctance of the narrow pole-tip structure of Fig. 1(a) is higher than that of the broad pole-tip structures of Figs. 1(b) and 1(c). Therefore the slope of the closed-circuit demagnetization line OB will be less in the former case and the flux threading the coil for a given magnetization of the medium will also be less.

The open-circuit signal voltage at the reproducing head is proportional to the rate of change of flux threading the coil. This factor in itself is responsible for a 6 db per octave rise in the signal characteristic. There are, however, several other factors that enter into the overall frequency response. (1) In both recording and reproduction the eddy-current characteristics of the head may introduce some attenuation at higher frequencies. (2) Following recording, the medium demagnetizes to a degree which is a function of the length of the recorded magnets. In longitudinal recording the resultant attenuation is an inverse function of the recorded wavelength. (3) In reproduction the response falls off when the dimension of the recorded wavelength becomes comparable to the active region of the reproducing head. Although the region in which a magnetized element of the medium may cause flux to thread the coil is not sharply defined, an effective dimension may usually be assigned, and this dimension has been called the "slit width" because of its similarity to the optical effect. In the ring-shape recording head experience has shown that the slit width is from 10 to 40 per cent greater than the air-gap length. (4) In certain reproducing heads, such as that of Fig. 1(b), the construction is such that the pole pieces are somewhat active over their entire dimension. Thus there is in effect a secondary slit width equivalent to the overall dimension of the poles which is responsible for irregularities in the response at longer wavelengths.

Lübeck has given, for ring-shape heads, a general expression for the open-circuit output voltage

$$E = CNI \frac{v}{s} e^{-\tau/\lambda} \sin \frac{\pi s}{\lambda} \tag{1}$$

where E and I are, respectively, the open-circuit output voltage and input current, C = a constant for the particular system, N = number of turns in reproducing coil, v = linear speed of medium, s = effective "slit width" of reproducing head, τ = a demagnetization constant for the medium, $\lambda = v/f$ = wavelength of recorded signal, f = frequency, and $\omega = 2\pi f$. The (electric to magnetic) frequency characteristics of the recording and reproducing head have been neglected in this expression.

Figure 5, curve (1), shows an illustrative frequency response characteristic. The factors determining the characteristic are independently plotted. It is seen that the function $\sin \pi s/\lambda$ which is the "slit width" characteristic plotted in curve (3) rises to a maximum

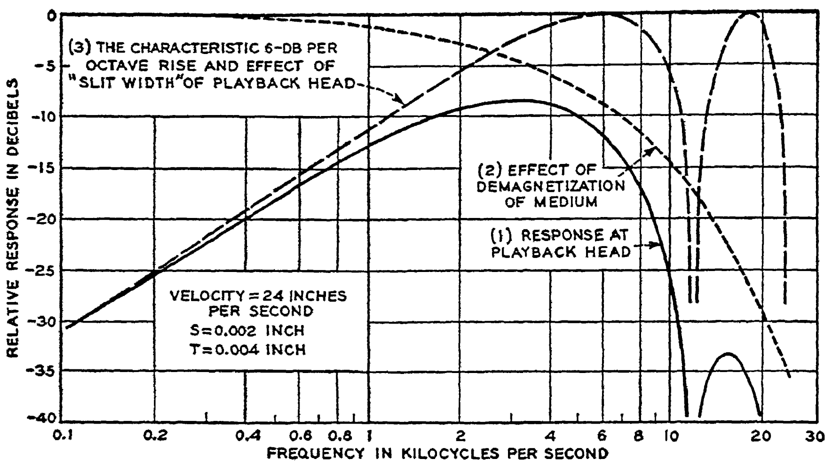


FIG. 5. Frequency Response of a Longitudinal Recording. Open-circuit playback voltage for constant recording current.

when the recorded wavelength is twice the effective slit width. Thereafter it goes through a series of dips and peaks. In a practical head design where the working gap is very small, the second and succeeding peaks are not reproduced because they lie in the region where the demagnetization, curve (2), of the medium causes a high loss and they usually fall outside the frequency range of associated electrical circuits.

At recording speeds in current use the slit width may be made sufficiently small so that $\sin \pi s/\lambda$ can be replaced by the angle itself within the working frequency range. Under these conditions

$$E = CNI \frac{v}{\lambda} e^{-\tau/\lambda} = CNI\omega e^{-\tau/\lambda} \tag{2}$$

and it is seen that, except for demagnetization, the output voltage rises 6 db per octave and is not influenced by the recording or reproducing speed or the width of the effective slit.

It is usually necessary to introduce an electrical equalization in the associated recording-reproducing circuits in order to obtain the desired overall frequency response. A major portion of the equalization is assigned to the reproducing circuits where it is limited only by the frequency spectrum of background noise.

When due consideration is given to the overload characteristic of the recording medium and the energy distribution of the recording signal, some of the equalization may be introduced in the recording circuits.

SOURCES OF NOISE. Background noise arises from a number of sources, many of which may be minimized in careful design. Among such sources may be listed amplifier hum, thermal noise, external noise pickup in the reproducing head, crosstalk transferred from a strong signal on adjacent turns of the medium when wound on the storage spool, mechanical irregularities on the surface of the recording medium which affect the degree of contact with the reproducing head, variations in the magnetic properties of the medium, and variations in the cross-sectional area of the medium. The remaining, and basic, noise source appears to be caused by a random distribution of small magnetic inhomogeneities. Some measurements at the reproducing head of noise from this source have indicated a fairly uniform frequency distribution of noise voltage per cycle. In most cases, after equalization, this background noise has a smooth character and so is not particularly disturbing. Basic noise is a function of the degree of magnetization of the medium. It is very low for a medium which has been thoroughly erased by a-c demagnetization, but it may be increased to a considerable extent by the presence of the a-c biasing flux. The noise level may be considerably increased by the use of a reproducing head in which there is some remanent magnetization, and the use of a d-c bias increases the noise markedly. Wherever the noise is decidedly affected by the presence of recording bias, there is detectable, back of the signal, a noise which rises and falls with the magnitude of the signal. A certain amount of such noise is not objectionable as it is partially masked by the signal.

22. RECORDING MEDIA

PHYSICAL FORMS. Most of the early magnetic recording equipment employed a solid homogeneous steel wire or tape as the recording medium. Recent refinements in the art of electroplating magnetic alloys and developments in the use of magnetic-powder-coated paper and plastic materials have removed many of the restrictions on the form of the recording medium. Although there are some technical differences between the media, the choice of physical form in general is determined by the desired application.

Magnetic wire is generally used wherever a very long playing time is required or a compact equipment is necessary. The wire diameter is limited only by the required breaking strength, and diameters between 0.004 in. and 0.006 in. are most common. Splicing is accomplished by tying a square knot which passes over a properly designed head without difficulty. A longitudinal recording head is usually employed, and the wire rides in a groove which is worn in the head. This method of recording is designed to distribute the recording signal about the circumference of the wire and serve as a protection against variation of the reproduced signal as the wire twists. There is, however, a considerable variation in the strength of the recorded signal around the circumference of the wire, particularly at short wavelengths, and it is fortunate that smooth round wire does not evidence a random twisting when properly used. The chief disadvantage of the use of wire is the tendency to uncoil, snarl, and kink whenever a free end becomes loose. For this reason it is advisable to mount the wire in a magazine containing the supply and take up reels and remove the magazine as a unit wherever record storage is required.

Magnetic tape, either solid or plated, is most useful for applications requiring a continuous record-reproduce-erase cycle. In such applications the tape is prepared in the form of an endless belt which is spliced, at the junction point, with a butt weld. Whenever a more extended playing time is required the tape is reeled like motion-picture film. Recording equipment has been commercially produced, using 0.002 in. by 0.050 in. tape and the offset recording and reproducing pole pieces of Fig. 1(*b*) in which an extended frequency response and dynamic range are realized at a comparatively low operating speed. The solid magnetic recording materials, both tape and wire, have the common disadvantage that accurate dimensional tolerances, smooth surface conditions, and uniformity of heat treatment must be maintained on large quantities of a material that is inherently very difficult to fabricate.

Coated paper and plastic recording media are applicable to all uses where requirements on playing time are not excessive and its low remanence can be tolerated. It has been used in the form of tape, sheets, and disks. The material is cheap to manufacture and may be cut and spliced by cementing. The material is usually 0.0015 to 0.0025 in. thick. In contrast to steel wire and tape, when a reel of the material is removed from a machine, there is little danger of freely uncoiling, a characteristic that is very desirable in the home recording field. When used with a properly designed ring-type recording-reproducing head a favorable frequency characteristic may be obtained with a comparatively low operating speed. Output levels are much lower than can be attained with solid wires and tapes, and care is required in shielding the reproducing head and in the design of reproducing amplifiers. Although the inherent dynamic range of the coated media is quite large, thermal and hum noise levels in reproducing equipment may limit the practical dynamic range.

Certain special applications of magnetic recording have employed magnetic media in the form of solid disks or cylinders. Most frequently the recording surface is electroplated on a non-ferrous backing material.

MAGNETIC MATERIALS. A carbon-steel recording medium was used in most of the early equipment because it is commercially available in the form of music wire. It has very little value in present-day recording applications. The coercive force is low, approximately 40 oersteds, and thus self-demagnetization is very pronounced. The residual induction is approximately 10,000 gausses. Because of demagnetization the short-wavelength (high-frequency) response from a longitudinal recording is very limited. The material magnetizes easily from the recording on adjacent turns of the storage spool, and crosstalk is therefore high. The basic background noise is rather high and is increased by surface corrosion of the material.

Thirteen per cent chrome steel has been used to a considerable extent in both magnetic wire and tape recorders because of its resistance to corrosion. This is a ferritic material depending mostly on the presence of carbon for its magnetic properties. The chromium is, however, instrumental in obtaining the desired hardness. When properly quenched, the material has a coercive force of 50–60 oersteds and a residual induction of 7000–10,000 gausses. A frequency response and dynamic range superior to carbon steel may be obtained from the material. High operating speeds are required, however, when an extended frequency range is desired. To prevent formation of chromium oxides, which are very abrasive, the material is heat treated in an atmosphere free of oxygen.

Nickel-chromium stainless steels such as 18 per cent chromium, 8 per cent nickel have been used in magnetic wire recorders. This is an austenitic steel which may be hardened by cold working. A somewhat more uniform product is obtained when the cold-worked material is then age hardened. After such treatment the coercive force is of the order of 150 to 350 oersteds with a possible maximum residual induction of 7000 gausses.

Several other magnetic alloys possess properties which make them exceptionally satisfactory as magnetic recording media. One such material is known as Vicalloy, a workable permanent-magnet alloy which is heat treated to obtain the desired magnetic properties. A typical Vicalloy composition is 38 per cent iron, 52 per cent cobalt, and 10 per cent vanadium. Vicalloy recording tape has been commercially produced with a coercive force of 225 oersteds and a residual induction of 6000 gausses.

One of the chief disadvantages of the solid magnetic recording media has been the difficulty of maintaining a uniform recording sensitivity and background-noise level throughout the entire length of the medium. Recently, an electroplated medium has been developed in order to correct this disadvantage and at the same time provide a cheaper recording material. The recording surface consists of a thin layer (approximately 0.0003 in.) of a nickel-cobalt alloy plated on hard brass wire or tape. The magnetic and physical properties are controlled in the plating process, and subsequent working or heat treatment is not required. The coercive force is approximately 200 oersteds with a residual induction of the order of 8000 gausses. A satisfactory frequency range may be obtained from this material at a comparatively low operating speed.

The use of powdered magnetic materials applied to paper or plastic carriers has recently received considerable attention in this country. Such a recording medium is comparatively cheap to manufacture and may be produced with very uniform and stable properties. Various magnetic materials in powder form are being investigated for properties advantageous to magnetic recording. One such material, black magnetic iron oxide, is commercially available in the required finely divided form. The powdered material is normally dispersed in a plasticized lacquer and applied to the carrier to a thickness of approximately 0.0005 in. When the powdered magnetic material is very finely divided and uniformly dispersed in the binder very excellent results have been obtained both in frequency response and background noise. Coercive forces ranging between 100 and 500 oersteds have been measured. The residual induction is very low, and of the order of a few hundred gausses. Demagnetization of signals of short wavelengths appears to be less pronounced in the powdered materials than in solid materials of equal coercive force. Output levels are very low, and, unless a very wide sound track is used, unusual precautions are required in the design of reproducing equipment which will take full advantage of the inherent dynamic range of the medium.

In Germany considerable work has been done on tape in which the magnetic medium is ferric and ferrous oxide in individual particles about 1 micron in size. This material is manufactured from precipitated finely divided black magnetic iron oxide by further oxidization in an agitated drier. The red ferric oxide has the crystal structure of the magnetic oxide and is also magnetic. Several types of recording tape have been manufactured in which the magnetic oxides are either cast on the surface of a plastic carrier such as cellulose acetate or polyvinyl chloride or are dispersed throughout a tape of polyvinyl chloride in a 50-50 mixture. It is reported that a frequency response uniform to within ±2 db from 50 to 10,000 cps is obtained at a tape speed of 30 in. per sec, the overall noise level is very low, and the useful life of the recorded tape exceptionally high.

BIBLIOGRAPHY

V. Poulsen, *Ann. d. Physik*, Vol. 3, 754 (Dec. 13, 1900).
V. Poulsen, *Electrician*, Vol. 46, 208 (Nov. 30, 1900).
L. Rellstab, *E.T.Z.*, Vol. 22, 57 (Jan. 17, 1901).
J. H. West, *E.T.Z.*, Vol. 22, 181 (Feb. 21, 1901).
A. Nasarischwily, *E.T.Z.*, Vol. 42, 1068 (Sept. 22, 1921).
W. L. Carlson and Glenn W. Carpenter, U. S. Pat. 1,640,881, applied for March 26, 1921; issued Aug. 30, 1927.
C. Stille, *E.T.Z.*, Vol. 51, 449 (March 27, 1930).
E. Meyer, *Handbuch der experimental Physik*, Vol. XVII, Chapter 11.
C. N. Hickman, *Bell Lab. Rec.*, Vol. 11, 308 (June 1933).
R. F. Mallina, *Bell Lab. Rec.*, Vol. 13, 200 (March 1935).
E. Schüller, *E.T.Z.*, Vol. 56, 1219 (Nov. 7, 1935).
H. Lübeck, *Akustiche Zeit.*, Vol. 2, 273 (November 1937).
C. N. Hickman, *Bell Sys. Tech. J.*, Vol. 16, 165 (April 1937).
S. J. Begun, *J. S.M.P.E.*, Vol. 29, 216 (August 1937).
A. E. Barrett and C. J. F. Tweed, *J. I.E.E.*, Vol. 82, 265 (March 1938).
D. E. Wooldridge, U. S. Pat. 2,235,132, applied for July 29, 1939; issued March 18, 1941.
M. Camras, U. S. Pat. 2,351,004, applied for Dec. 22, 1941; issued June 13, 1944.
S. J. Begun, *Proc. I.R.E.*, Vol. 29, 423 (August 1941).
Marvin Camras, *Radio News*, Radionics Dept., Vol. 1, 3 (November 1943).
Hershel Toomim and David Wildfeuer, *Proc. I.R.E.*, Vol. 32, 664 (November 1944).
L. C. Holmes and D. L. Clark, *Electronics*, Vol. 18, 126 (July 1945).
S. J. Begun, *Communications*, Vol. 26, 31 (April 1946).
D. E. Wooldridge, *Trans. A.I.E.E.*, Vol. 65, 343 (June 1946).
C. W. Hansell and others, Dept. of Commerce, Office of the Publication Board, PB 1346.
J. Z. Menard, Dept. of Commerce, Office of the Publication Board, PB 12659.

MECHANICAL RECORDING AND REPRODUCING OF SOUND

By L. Vieth and H. A. Henning

The requirements for high-fidelity recording and reproduction of sound are (*a*) that the whole system, from the point where the sound reaches the pickup device to the point where it is actually reproduced as sound, shall have a linear relationship between its input and its output, (*b*) that the system have a uniform response versus frequency characteristic, since any complex wave may be resolved into the sum of simple sinusoidal terms, and (*c*) that the system have a linear relationship between its phase shift and the frequency impressed upon it and that the phase angle have a value of $\pm n\pi$ at zero frequency where $n = 0, 1, 2, 3$, etc. The requirements of individual components may vary from the requirements of uniform response versus frequency characteristic (*b*), but in general the requirements for the system apply to its components.

Most sound recording and reproducing systems and components represent a compromise between the requirements for high-quality reproduction and the commercial requirements of size, cost, and general adaptability for a specific use. These compromises have resulted in a wide range of overall performances. Space limitations prohibit discussion of more than a few of the typical instruments in commercial use in the mechanical recording and reproducing of sound.

23. RECORDING INSTRUMENTS

Mechanical recording is used almost exclusively in the present-day phonograph industry, in electrical transcriptions for broadcast purposes, and in stenographic applications (dictation machines) and general utility recorders. Most mechanical recording is done on disks varying in diameter from 6 in. for general utility equipment to 16 in. in broadcast transcriptions. There are, in addition, several commercial devices in which the recording medium takes the form of standard motion-picture film base. The materials of which these media are made are discussed elsewhere. They may be grouped, in general, into two classes: (*a*) materials for recording for instantaneous playback; (*b*) materials for recording for subsequent processing. However, the function of the recording instrument is the same in all: to transfer to the recording medium, in the form of embossed or engraved modulations, a counterpart of the voltage impressed on the recorder modified in frequency characteristic in conformance with some preconceived plan of equalization which affords the optimum use of the recording equipment.

A typical recording instrument used in this work is illustrated by Fig. 1, which shows a phantom view of a lateral-type recorder (one in which the stylus moves laterally parallel

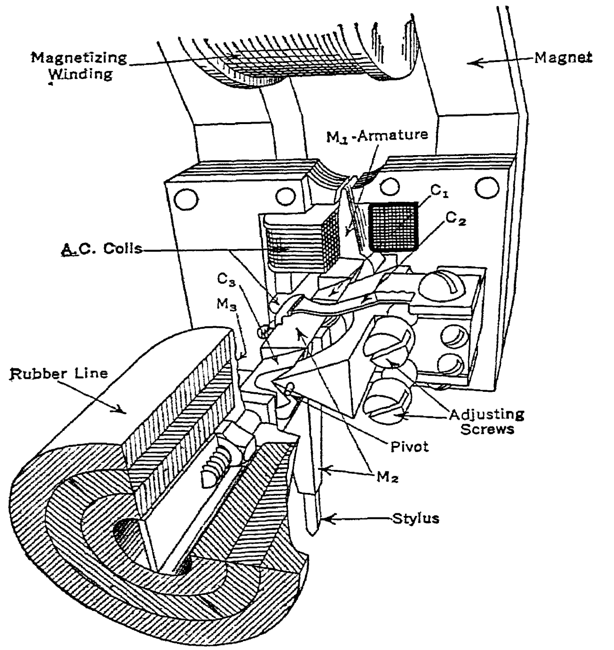


FIG. 1. Electromechanical Type Recorder

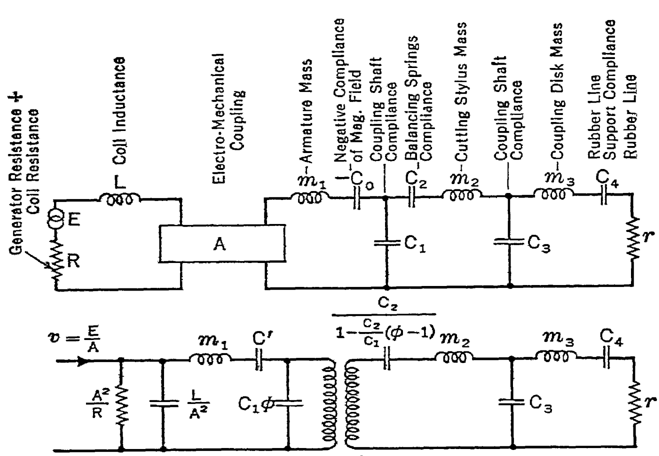


FIG. 2. Equivalent Circuits of Electromechanical Recorder Shown in Fig. 1

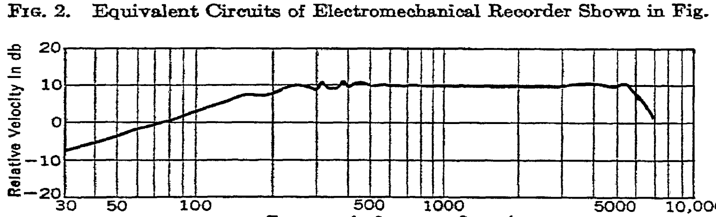


FIG. 3. Response vs. Frequency Characteristic of a Lateral Type Recorder

to the radius of the disk) designed by Bell Telephone Laboratories some years ago. The same general structure has been used in vertical-type recorders (one in which the stylus moves normal to the face of the disk). The structure is a mechanical filter whose electrical equivalent is shown in Fig. 2, in which the current in the second mesh is analogous to the

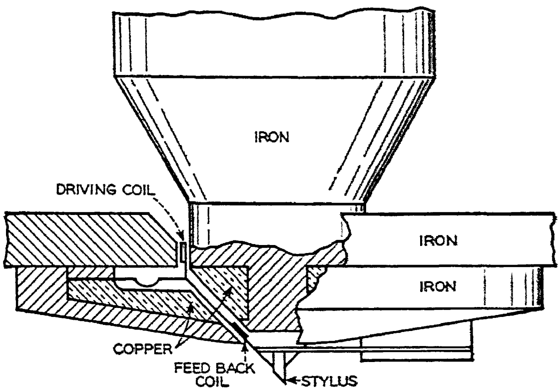


FIG. 4. Cross-sectional View of Vibrating System and Associated Magnetic Circuit of Western Electric Co. 1A Feedback Recorder

stylus velocity. Figure 3 shows a typical response versus frequency characteristic of such a recorder. The loss at the lower frequencies is part of a preconceived equalization plan which makes the best use of the recording medium by limiting amplitudes in the interest of record space economy.

The vibratory system of a more modern version of recording instrument is shown in Fig. 4. This device, also designed at Bell Telephone Laboratories, utilizes the principle of stabilized feedback to control the stylus velocity and involves an associated amplifier in which the recorder becomes an actual transmission element as well as a terminal transducer. Schematically, the device associated with an appropriate amplifier may be represented as shown in Fig. 5. The output voltage E_2 of the amplifier is supplied to the driving coil of the recorder, thereby driving the stylus with a velocity V. Motion of the stylus in turn generates in a suitable generating element, such as a small coil moving in a magnetic field, the voltage E_3. This voltage is returned to the amplifier input through a control circuit which may be either passive or active. The voltage available after modification in the control circuit is designated E_4.

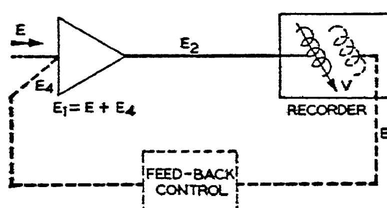


FIG. 5. Schematic Representation of an Electromechanical Feedback System

The voltages and velocities here referred to are to be considered as having both magnitude and phase and hence can be represented in complex number notation. Then

$$E_1 = E + E_4 \tag{1}$$

To obtain a simple expression for the relation of the stylus velocity V to the signal voltage E, let

$$A = \frac{V}{E_1} = \frac{E_2}{E_1}\frac{V}{E_2} \tag{2}$$

and

$$B = \frac{E_4}{V} = \frac{E_3}{V}\frac{E_4}{E_3} \tag{3}$$

and hence

$$AB = \frac{E_4}{E_1} \tag{4}$$

The product AB thus defines the transmission around the loop formed by the amplifier, recorder, and feedback control. The value of E_4 from this equation can now be substituted in the relation $E_1 = E + E_4$ to obtain

$$E_1 = E \frac{1}{1 - AB} \tag{5}$$

which, together with eq. (2), gives

$$V = E_1 A = E \frac{A}{1 - AB} \tag{6}$$

The right-hand side of eq. (6) is the familiar expression for feedback amplifiers in general, and the rules for stability, gain, distortion, etc., are equally applicable. In particular, when AB is very large compared to unity

$$V = \frac{1}{-B} E \qquad (|AB| \gg 1) \tag{7}$$

which indicates that over the frequency range considered the velocity of the stylus is independent of the amplifier gain or the efficiency of the recorder. Variations in B, however, directly affect the performance, and hence, if a flat frequency response is desired, B must remain constant. However, since B is the product of the mechanical-electrical conversion factor E_3/V and the control factor E_4/E_3, it will be seen that these factors may vary as long as their product remains constant. It is a simple matter to maintain the factor E_3/V constant, and hence a flat response characteristic depends only upon keeping the control factor constant.

If eq. (6) is rewritten to include noise and distortion products as well as signal, it becomes

$$V = E \frac{A}{1 - AB} + \frac{n}{1 - AB} + \frac{d}{1 - AB} \tag{8}$$

where n and d are the noise and distortion, respectively, introduced in the amplifier and recorder without feedback. Hence, when AB is large compared to unity, both the noise and the distortion components are reduced as compared with the corresponding effects in a non-feedback system.

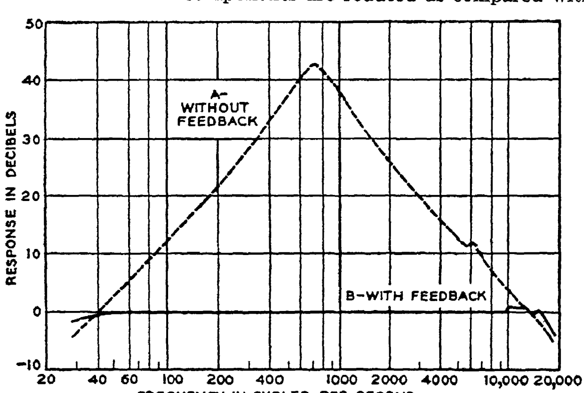


Fig. 6. Curves Showing the Stylus Velocity for a Constant Signal Input to the Recorder Amplifier System. (A) Without feedback; (B) with properly controlled feedback.

Variations in the impedance of the recording medium which act upon the stylus during cutting may be regarded as noise or distortion introduced in the recorder, and their effect upon the vibrational velocity is also reduced by the above factor. This is equivalent to a manifold increase in the driving-point impedance. A frequency response characteristic with and without feedback on a typical feedback-type recorder is shown in Fig. 6. The flat frequency response characteristic shown is, of course, of limited use in disk recording since the large amplitudes at low frequencies are prohibitive. However, since the overall characteristic of the device is flat, pre-equalization can be accomplished flexibly and simply by electrical networks.

24. RECORDING AND REPRODUCING MEDIA

Two broad classifications may be made of mechanical recording media: (a) the "wax" disk intended primarily for subsequent processing and duplication, and (b) the instantaneous media, disk or otherwise, intended mainly for reproduction directly from the embossed or engraved master. The latter are occasionally used for subsequent processing and duplication. "Wax" cylinders are used to some extent in dictating machines as direct playback media.

RECORDING DISKS FOR PROCESSING AND DUPLICATION. The shape of the recording stylus varies somewhat in commercial practice in the cutting of disks for subsequent processing and duplication. In general the cross-section and particularly the curvature of the unmodulated groove are kept within fairly narrow limits. The stylus used on Western Electric recorders has a tip radius of 0.0021 in. ±0.0001 in. Its contour is illustrated in Fig. 7. The disposition of grooves and land between grooves is a matter of compromise between playing time and signal-to-noise ratio and other factors. For lateral recording the groove pitch is usually between 0.007 and 0.010 in. as illustrated by Fig. 7. The maximum safe amplitude at any point on the groove will depend upon the amplitude and phase of the sound recorded on the adjacent grooves. For a groove situated between two unmodulated grooves, the maximum amplitude for the condition pictured in Fig. 7 will be nearly 0.004 in. The maximum amplitude for two adjacent grooves of equal amplitude and 180° out of phase with each other will be slightly less than 0.002 in. The space between grooves may be compressed, in vertically cut recordings, to increase the playing time of a record with less sacrifice in recording level than for lateral recording.

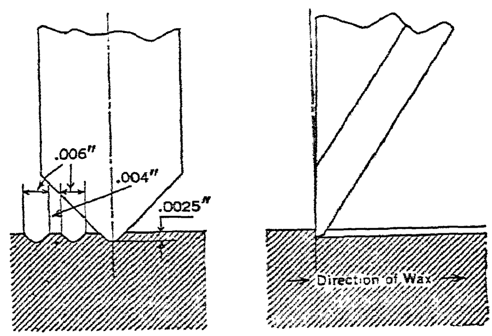


FIG. 7. Western Electric Co. Lateral Recording Stylus

In cutting, the "wax" (more correctly a metallic soap) must be leveled on the recording machine with reasonable care, and the stylus must be sharp and so ground that the cut will be very clean. As it is cut the wax shaving is removed by air suction. The operator is aided in maintaining the correct depth of cut by a so-called advance ball. Commonly an advance ball is a ball-shaped sapphire mounted in an adjustable holder which is in turn fastened to the recorder. The advance ball rides lightly over the wax close to the stylus and serves to maintain uniform depth of cut in spite of small inaccuracies of leveling of the "wax" or deviation from planeness. The advance ball is adjusted relative to the stylus by observing the cut with a calibrated microscope.

In commercial cutting, for processing, both solid and flowed waxes are popular. Solid disks are shaved with a sapphire knife to a highly polished surface on a sturdy high-speed turntable. After recording and processing the disk is reshaved and recut until it becomes too thin to be used with safety. The flowed disk is a thin layer of wax flowed onto a metal surface. The wax layer is cut once and discarded, the metal backing being reused indefinitely. Such a recording medium is extremely smooth, homogeneous, and free from mechanical strains incidental to shaving the "waxes" by the older method.

DUPLICATION OF DISK RECORDS. The surface of the wax, after being engraved, is rendered electrically conducting. This can be done in a number of ways, such as dusting with graphite and bronze or other electrically conducting powders, by the chemical precipitation of silver, or by sputtering a suitable material such as gold or silver in a cathode sputtering chamber.

The wax is then electroplated with copper, sometimes to a thickness of about $^1/_{32}$ in., or sometimes only to a thickness of a few thousandths of an inch and then backed up by a thicker metal plate of suitable material. This electrodeposited plate is then a negative of the wax and is called a "master." From this negative or master, positive copies may be made. In commercial practice the copies or records are usually pressed of some thermoplastic material. If the sound reproduction from a few records carefully pressed from a master is satisfactory, duplicate stampers are produced in order that the master may be preserved. These duplicate stampers are used for pressing or molding the commercial release records.

From the master one or more "mother" plates are made by electrodeposition. These mothers are positives and serve as the forms on which the final matrices or stampers are deposited. One of the important problems in this process is the treating of the metal surfaces of the masters and mothers so as to permit subsequent separation. The problem is to produce a surface which will still be conducting, but which is not "clean"; i.e., the deposited metal must not be in sufficiently intimate contact with the metal of the mold to cause permanent adherence. This may be done by either of two methods: (*a*) by the application of a thin "mechanical" layer of some substance such as grease, graphite, etc., or (*b*) by a chemical film usually produced by treating the previously cleaned metal surface with some chemical that will react with the metal and produce an insoluble compound. Typical substances for the purpose are soluble sulfides such as sodium sulfide (Na_2S) or the yellow "polysulfide" (Na_2S_x). These will react with the metal to form films of the corresponding sulfides. Frequently the phonograph industry has employed a film of silver iodide, formed by first treating the copper surface with a silver cyanide solution. The solution is either poured over the previously cleaned surface or mixed with an inert substance such as whiting or calcium carbonate ($CaCO_3$) to form a paste which is rubbed over the copper. A thin film is thus deposited by immersion. The film is then treated with a weak iodine solution. The iodine solution acts upon the silver to form a thin film of silver iodide upon which the copper can then be deposited and subsequently separated. Numerous other methods are employed in the industry.

Until recently the records themselves were molded of a thermoplastic mixture of shellac and earth fillers and were generally played with steel needles. The fillers were purposely made somewhat abrasive in order to grind the needle point to fit the groove since the needles were not usually accurately shaped to a suitable contour. The pressures at the needle point of a typical needle, before being ground to shape, are extremely high, and record wear under this condition is serious. In more recent years, reproducers of lower mechanical impedance have become available, which require much lower pressures. Accurately contoured permanent reproducing points provide further relief by reducing vibratory mass and maintaining a definite small radius of curvature. Any abrasive in the records adds to the background noise, hence homogeneous, non-abrasive records such as those made of cellulose acetate or vinylite are coming to be used. Such materials are, in fact, used almost universally in transcription recordings for broadcast purposes.

Records of the older shellac mixtures or of some of the more recently developed plastic materials are made in about the same manner, although the time, temperature, and pressure cycles may differ widely. The material to be molded is usually heated to a suitable softening temperature (of the order of 300 deg fahr) either on an auxiliary hot plate or in the record press itself in which the stampers are mounted. The stampers are commonly mounted in the press on platens which are heated and then cooled according to some predetermined cycle. The material may be further heated in the press at a suitable temperature before the pressure of the press is applied upon it. The pressure applied is of the order of 2000 lb per sq in. for many plastic materials. This pressure is maintained while the platens are cooled for a suitable time, after which the press is opened and the record removed.

INSTANTANEOUS PLAYBACK DISKS. This term covers a multitude of plastics and metals which lend themselves to easy engraving or embossing. A few will be discussed. In the higher-quality field of instantaneous recording media the 12 in. and 16 in. lacquer-coated disks of glass or metal are most common. A variety of materials may be compounded with cellulose nitrate to form a lacquer that is easily engraved. These coatings are approximately 0.010 in. thick and cut cleanly with a recording stylus similar to that used for cutting "wax," although it is general practice to somewhat dull the cutting edge to a radius of the order of 0.0003 in. The shaving may be directed toward the center and collected about the center pin of the disk or it may be removed by suction apparatus.

Disks of this kind are sometimes used as masters for subsequent processing and duplication in the same manner as described for wax disks. When the disks are used for instantaneous playback, the elastic properties of the recording material seriously affect the response characteristic. High-frequency losses, particularly at low linear velocities, are very severe and are not too satisfactorily relieved by predistortion in recording. Depending upon the type of reproducer used on these disks, they may be good for from one to fifty playings.

In the general utility field recording disks are most commonly made of vinylite or vinylchloride. Gelatin and soft metals such as aluminum are also used. These materials, in general, are not suitable for engraving, and in most applications an embossing stylus is used which rubs a shallow groove into the recording material, modulating the groove in accordance with the signal impressed on the recorder. The contour of such styli varies widely. Grooves so rubbed are about 0.001 in. deep and the maximum amplitude rarely exceeds 0.001 in. Signal-to-noise ratios are relatively low. These applications are definitely not high quality in their present state of development, emphasis being placed on obtaining good articulation. Useful playing life is uncertain. Extremely light reproducers and soft reproducing styli are required if more than a few playings are anticipated. Disks in general do not exceed 6 or 8 in. in diameter.

RECORDING MEDIA OTHER THAN DISK. Although the recording medium on most general-utility recording machines takes the form of a disc, two exceptions worthy of note are made. The first of these is film on which the recorded grooves are embossed in helical pattern on an endless loop of standard 35-mm safety motion-picture film base

(Amertype Recordgraph). The other exception is the well-known wax cylinder which approximates in its consistency the wax disk used in recordings for processing. Here a helix is cut in the cylinder to be erased for subsequent reuse after playing (Dictaphone and Ediphone).

A unique combination of mechanical recording and optical reproduction is found in the Philipps-Miller system, in which a wide-angle stylus cuts an extremely shallow groove through the darkened emulsion of a photographic film. The thin emulsion is therefore cut away in varying widths so that the transparent body of the film presents a pattern similar to that of variable area sound-on-film recording. Reproduction is accomplished optically in a similar manner to ordinary sound on film.

25. REPRODUCING INSTRUMENTS

The function of the reproducer in a mechanical recording system is to transform into electrical energy the modulations of the recorded disk or duplicate thereof. Since the inception of so-called "electrical recording" in contrast to the older method of acoustical recording, magnetic-type reproducers have played an important part. Such reproducers have taken a variety of forms, but in general an armature is rigidly attached to a stylus holder. The armature, moving in a magnetic field, generates a voltage in associated coils. Typical examples of such structures are illustrated in Figs. 8 and 9. Both these devices and others of the same vintage are lateral-type reproducers and are intended for use with a replaceable steel stylus or needle. To provide sufficient rigidity against flexing, such styli in themselves are quite heavy, and their mass plus that of associated driving elements produces a

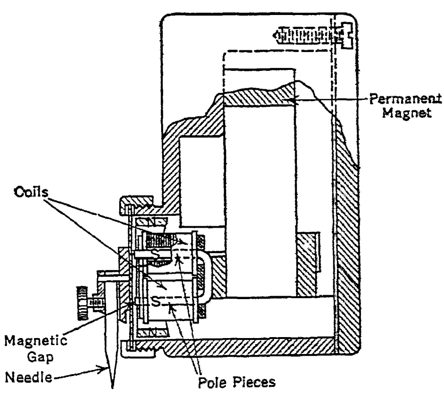


FIG. 8. Early Type Western Electric Co. Oil-damped Lateral Reproducer

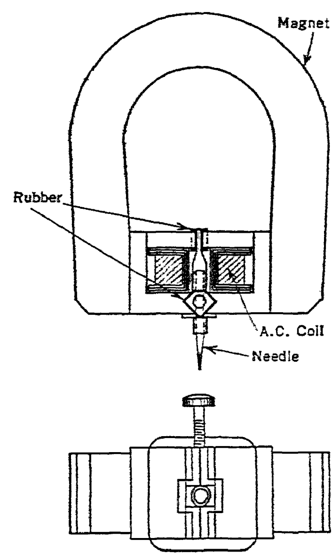


FIG. 9. Early Type RCA Victor Reproducer

high driving-point impedance with corresponding record wear and distortion. In general, the same applies also to the fitted jewel styli used in conventional-type lateral reproducers.

In reproducing vertical recordings the driving force is in direct line with the stylus axis and flexural rigidity is not so important a requirement. The mass of the stylus and associated vibratory system can therefore be kept comparatively small, thus reducing the driving-point impedance so that record wear is greatly reduced. Vertical reproducers in general have been built around the moving-coil principle and invariably are associated with a jeweled (diamond or sapphire) stylus that has been polished to fit the recorded groove accurately.

One of the more common of reproducers is the piezoelectric crystal type, in which a Rochelle salt crystal element is secured to a stylus holder at one end and anchored at the other. Stylus movement causes a flexing of the crystal element, which produces a voltage. In general that type of reproducer is of fairly high driving-point impedance, although some of the higher-quality devices provide a distinct improvement in this respect. Figure 10 shows a cross-section view of the "cartridge" of a typical crystal-type reproducer.

An interesting reproducer which serves to illustrate at once the moving-coil principle

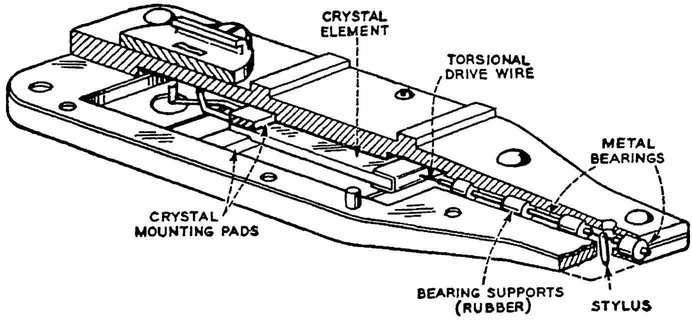


FIG. 10. Brush Development Co. PL-20 Crystal Reproducer

as applied to both vertical- and lateral-type reproducers is the Western Electric 9A reproducer, a phantom view of which is shown in Fig. 11. This is a universal reproducer, that is, one which will reproduce either vertical or lateral recordings by changing the electrical

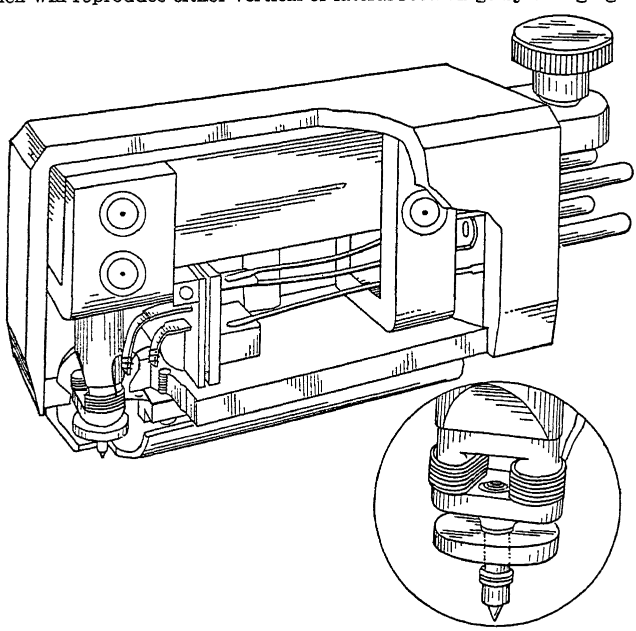


FIG. 11. Western Electric Co. 9A Universal Reproducer

relationship of the two voltage generating coils. Figure 12 shows the response characteristic of this device on both types of recordings.

Jeweled styli have become commonplace in many makes of reproducers. Their contours vary, but in general the radius of the spherical tip is held to between 0.002 and 0.003 in. The larger tip radii, though permissible and even desirable in certain types of lateral reproducers, in general induce serious distortion products.

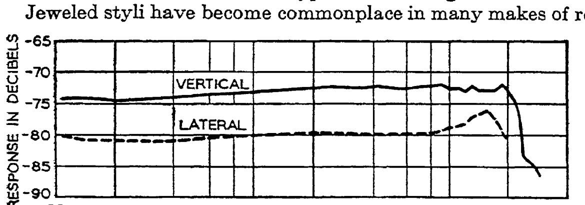


FIG. 12. Response vs. Frequency Characteristic of Typical 9A Reproducer

They do, of course, reduce wear on the record.

26. SOURCES OF DISTORTION

RECORDING. In all recording instruments sources of distortion common to most electromechanical transducers are present. Magnetic elements do not always behave linearly, and mechanical elements are frequently free to respond in modes other than the desired one. In general, however, distortion due to these effects can be minimized by good engineering and design. Recorders intended for use in a variety of recording media do not always meet the same impedance at the cutting point. Not only does this impedance vary from one medium to another; it may vary also with depth of cut in a given medium. It has been customary, therefore, to keep the driving-point impedance sufficiently high so that the recording medium has little or no effect on the recorder characteristics. The maximum force on the stylus in cutting is developed by the reaction of the record material to be cut away. This force is normal to the motion of the stylus, and yielding by the stylus arm in this direction occurs in the neighborhood of the stylus-arm resonance frequency. This motion is introduced on the time axis of the recording, and intermodulation with other frequencies results.

A less important source of distortion which intrudes also on the smoothness of cutting is the design of the stylus. The dimensions of the stylus, like so many other elements in recording systems, are a compromise between many factors, including linear velocity and recording amplitude. In order to provide adequate ruggedness at the tip, a short bevel at a rather blunt angle is provided, as is shown in Fig. 7. The slope of this bevel marks the maximum slope that can be tolerated in vertical recording without having the heel of the stylus abrade the modulated groove. The corresponding critical angle in lateral recording is approximately 45°, which means that the maximum recorded vibratory velocity must never be more than the linear velocity of the record.

PROCESSING AND DUPLICATION. Although electrodeposition is one of the most accurate methods known for duplicating a surface, several sources of distortion present themselves. Plating on the face of copper stampers of abrasion-resistant chromium or other metallic finish has a finite thickness sometimes comparable with the modulations of the grooves. When this process is repeated a number of times, as is common commercial practice, high frequencies are all but obliterated. Stampers are occasionally buffed before or after plating. This, of course, very rapidly erases, in a random manner, much of the higher frequencies.

REPRODUCING. The distortions introduced in recording and processing can be minimized by good design and careful handling. In the interest of commercial expediency these factors are sometimes overlooked. In reproducers, however, there are in the present state of the art several distortion-producing factors which might be characterized as inherent. The first of these, termed "tracing" distortion, is due to the fact that the reproducer stylus has finite size. The curve traced by the center of a spherical-tip stylus in tracing a sinusoidally modulated groove is not sinusoidal. This effect increases rapidly as the minimum radius of curvature of the recorded wave approaches the radius of the stylus tip. Practical considerations prevent relief by the simple expedient of making the stylus radius smaller.

The minimum radius of curvature of a sine-wave groove is given by

$$R = \frac{V^2}{100\pi^2 A f^2} \tag{9}$$

where R = minimum radius of curvature (inches), V = linear speed of groove (feet per minute), A = amplitude (inches), and f = frequency (cycles per second). The linear speed V of a 12-in. record at the standard turntable speed for phonographs, 78.26 rpm, varies between 1.3 and 3.8 ft per sec. At the standard broadcast transcription speed of 33.33 rpm the linear speed of a 16-in. record varies between 1.2 and 2.4 ft per sec. At high frequencies, therefore, for even extremely small amplitudes the radius of curvature of the modulated groove may be comparable with that of the stylus attempting to trace it. Fortunately, the distribution of energy in speech and music spectra is such as to alleviate this situation. However, in the interest of noise reduction and standardization of recording techniques it is an accepted procedure to accentuate high frequencies in recording with corresponding attenuations in reproduction.

In lateral reproduction the problem is further complicated by what has been described as "pinch" effect; i.e., since conventional lateral reproducers have no vertical compliance the stylus "pinches" between the walls of the grooves when the linear velocity drops below a certain point. This phenomenon results from the fact that the cutting surface of the recording stylus is always perpendicular to the unmodulated groove. A constriction therefore results in the width of those portions of a modulated groove which are at an angle to the direction of the unmodulated groove, i.e., at all portions of the modulated grooves except the maxima and minima and points of inflection. Some relief is afforded by providing appropriate vertical compliance in lateral reproducers, and in such reproducers it is advantageous to use an oversize reproducing stylus which never rides on the groove bottom but is always positively driven by the side walls. It has been shown that in this way even harmonics can be eliminated from "tracing" distortion.

Another source of distortion in reproduction is due to "tracking error." Tracking error may be defined as the angle between the vibration axis of the mechanical system and the tangent to the groove being reproduced. This angle results from the conventional device of pivoting the reproducer arm at a fixed point. The aforementioned vibration axis can therefore be truly tangent to the record groove at only one radius. In lateral reproduction when the tracking error is large a sinusoidal wave is not traced sinusoidally. This effect is minimized in lateral reproduction by the well-known device of an offset reproducer head or by making the pivoted reproducer arm very long. Mechanisms which keep the reproducer tangent at all times are employed in a few reproducing devices, eliminating this source of distortion entirely. In vertical reproduction, distortion due to tracking error is negligible.

The characteristics of the medium of which disks are made contribute to the distortion in reproduction. The modulated groove which drives a reproducer stylus has a finite impedance which at some frequencies may be comparable with or less than the stylus-point impedance. With modern high-quality reproducers the compliance of the record material resonates with the vibratory mass of the reproducer, frequently producing a peaked response at some high frequency, above which the response very rapidly declines. The most outstanding examples of the effect of record characteristics on response are found in the instantaneous types of playback disk. The materials of which these disks are made have a high compliance. The losses at high frequencies, particularly at low record velocities, are serious even with the best of commercial reproducers. Compensation for such losses is difficult because the magnitude of the loss varies widely, increasing as the linear velocity decreases.

BIBLIOGRAPHY

T. A. Edison, U. S. Pat. 200,521.
H. M. Stoller, *J. S.M.P.E.*, 1928.
O. J. Zobel, *B.S.T.J.*, 1928.
S. P. Mead, *B.S.T.J.*, 1928.
H. A. Frederick, *J. A.S.A.*, 1931.
E. C. Wente and A. L. Thuras, *J. A.S.A.*, 1931.
W. C. Jones and L. W. Giles, *J. S.M.P.E.*, 1931.
W. C. Jones, *J. S.M.P.E.*, 1931.
E. C. Wente, *Phys. Rev.*, 1922.
H. F. Olsen, *J. A.S.A.*, 1931.
H. C. Harrison and P. B. Flanders, *B.S.T.J.*, 1932.
H. D. Arnold and I. B. Grandall, *Phys. Rev.*, 1917.
Tucker and Paris, *Phil. Trans. Roy. Soc.*, 1921.
A. V. Hippel, *Ann. d. Physik*, 1925.
W. Spaeth, *Zeit. tech. Phys.*, 1925.
U. S. Patent 1,634,210, 1927.
Olson and Massa, *Applied Acoustics.* P. Blakiston's Son & Co., Philadelphia.
J. P. Maxfield and H. C. Harrison, *Trans. A.I.E.E.*, 1926.
C. F. Eyring, *J. A.S.A.*, 1930, 1931; *J. Soc. Mot. Pict. Eng.*, 1930.
R. L. Hanson, *J. S.M.P.E.*, 1930.
F. L. Hunt, *J. A.S.A.*, 1931.
W. A. MacNair, *J. A.S.A.*, 1930; *Proc. I.R.E.*, 1931.
J. C. Steinberg, *J. A.S.A.*, 1929.
W. C. Sabine, *Collected Papers on Acoustics.*
L. J. Sivian, H. K. Dunn, and S. D. White, *J.A.S.A.*, 1931.
J. P. Maxfield, *J. S.M.P.E.*, 1930.
L. Cowan, *Recording Sound for Motion Pictures.* McGraw-Hill, Chapters 4, 16.
E. O. Scriven, *J. S.M.P.E.*, 1928; *B.S.T.J.*, 1934.
W. P. Dutton and S. Read, Jr., *J. S.M.P.E.*, 1931.
S. Read, Jr., *J. S.M.P.E.*, 1933.
R. A. Miller and H. Pfannenstiehl, *J. S.M.P.E.*, 1932.
H. A. Frederick, *J. S.M.P.E.*, 1928, 1932.
P. Wilson and G. W. Webb, *Modern Gramophones and Electric Reproducers.* Cassell & Co. (1929).
S. T. Williams, *J. Franklin Inst.*, 1926.
H. A. Frederick and H. C. Harrison, *Trans. A.I.E.E.*, 1933.
A. Güntherschuze, *Zeit. f. Phys.*, 1926.
H. F. Fruth, *B.S.T.J.*, 1932.
W. Blum and G. B. Hogaboom, *Principles of Electroplating and Electroforming.* McGraw-Hill.
F. C. Barton, *J. S.M.P.E.*, 1934.
L. F. Rahm, *Plastic Molding*, McGraw-Hill (1933).
E. W. Kellogg, *Trans. A.I.E.E.*, 1927.
J. P. Maxfield, U. S. Pat. RE 18,228.
E. Nesper, *Nimm Schallplatten selver auf.* Franckh'sche Verlagshandlung, Stuttgart.
U. S. Pat. 1,754,935 (1930).
U. S. Pat. 1,421,045 (1922).
A. C. Keller, *J. A.S.A.*, Vol. 8 (1937).

A. C. Keller, *J. S.M.P.E.*, 1937.
M. J. DiToro, *J. S.M.P.E.*, Nov. 1937.
B. Olney, *Electronics*, November 1937.
L. Vieth and C. F. Wiebusch, *J. S.M.P.E.*, January 1938.
J. A. Pierce and F. L. Hunt, *J. S.M.P.E.*, August 1938.
H. A. Henning, *Bell Lab. Rec.*, October 1940.
W. D. Lewis and F. L. Hunt, *J. A.S.A.*, January 1941.
L. Fleming, *J. A.S.A.*, January 1941.
B. B. Bauer, *Electronics*, March 1945.
B. B. Bauer, *J. A.S.A.*, April 1945.
H. E. Roys, *J. S.M.P.E.*, June 1945.
W. S. Bachman, *Electronic Industries*, July 1945.

PHOTOGRAPHIC SOUND RECORDING

By C. R. Keith

The photographic method of sound recording finds its greatest field in sound pictures, the sound record being usually on the same film as the picture or at times on a separate film operating synchronously with the picture film. Film records are particularly adapted to sound pictures since these films are usually made from numerous short "takes" which are edited and then spliced together. The sound and picture are usually photographed on separate films, except for newsreel recording where the sound and picture are commonly taken on the same film. Usually music and sound effects are recorded separately from the dialog; the various sound tracks are then rerecorded in synchronism with the finally cut picture film to make a sound negative which is used for making release prints.

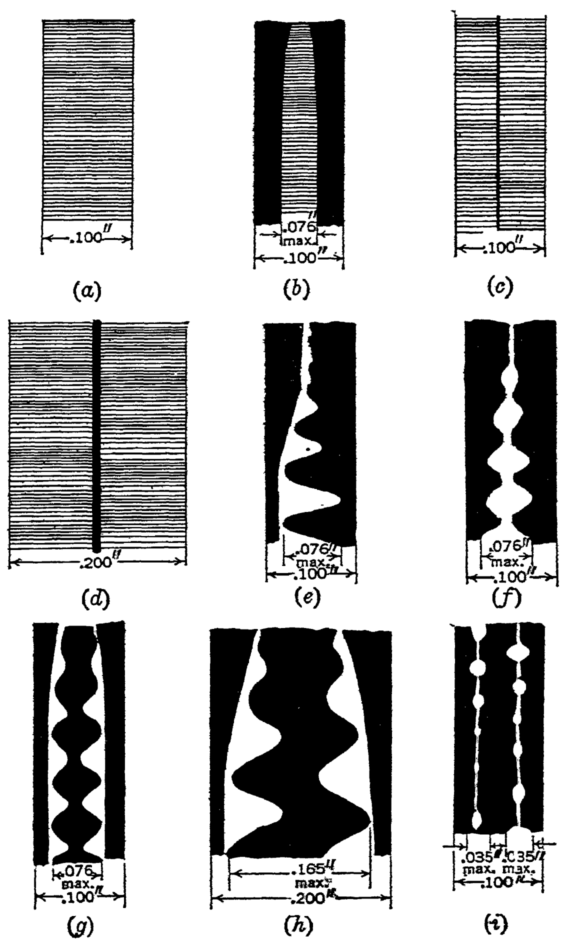


FIG. 1. Types of Sound Tracks

Two types of sound-on-film records are in common use today. One is the variable-density type—a series of striated bands as shown in Fig. 1 (*a*)–(*d*). The other is the variable-area type shown in Fig. 1 (*e*)–(*i*)—a serrated band with its toothlike projections. Both place the record on a narrow strip of the film at one side of the picture as in the sketch of a composite print with a variable-density sound track (Fig. 2). As the sound track must be played at uniform speed while the picture progresses with intermittent motion, it is displaced forward along the film 15 in. from the corresponding picture, so that the momentary difference in film velocity can be taken up by a free loop. Both types of sound track can be reproduced in the same machine without any change in the machine or in the electrical equipment.

The most commonly used sound tracks are illustrated in Fig. 1. Films exhibited in theaters ordinarily have single 100-mil tracks such as (*a*), (*b*), (*e*), (*f*), or (*g*). Original sound records are often made with 100-mil pushpull tracks such as (*c*) or (*i*) or more commonly, when the sound is recorded on a film separate from the picture film, with a 200-mil pushpull track such as (*d*) or (*h*).

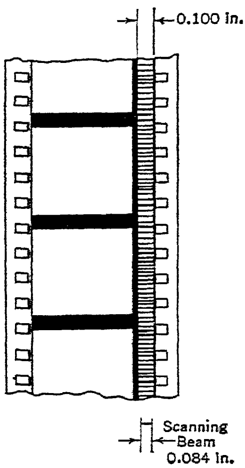


Fig. 2. Position of Sound Track on 35-mm Film

There are many methods of producing such film sound tracks: the light-valve method (Western Electric), the reflecting-galvanometer method (RCA), the flashing-lamp method (Fox-Case), the Kerr cell (Klangfilm), and many variations of these methods. The flashing lamp and the Kerr cell are used only for variable-density tracks, but the light valve and the reflecting galvanometer may be used for either variable-density or variable-area tracks.

27. LIGHT-VALVE RECORDING SYSTEM

In the light-valve system the light from an incandescent source is made to fall on a light valve, Fig. 3, formed of two strips of Duralumin 0.0005 in. thick by 0.006 in. wide spaced 0.001 in. apart. These are placed in a magnetic field and carry the speech currents. Since the two ribbons carry current in opposite directions they move together or apart as the current varies. In commercial practice the ribbons are tuned or resonated to about 9500 cps. The response in the frequency range near resonance depends on the damping, and in early types of light valves the response at resonance was 20 to 25 db higher than at low frequencies. This variation in response can be reduced to 1 or 2 db by using a feedback circuit in which signal voltage across the light-valve ribbons is amplified and applied to the valve input in opposite phase. However, the more recent light valves have sufficiently high flux density (30,000 gausses) so that the resonance peak is only about 6 db. Additional damping is obtained by passive networks in the valve input circuit, so that the resonance peak may be reduced to about 1 db. These measures also almost entirely eliminate extraneous vibrations of the light-valve ribbons when excited by steep transients.

In early types of light valves both ribbons were in the same plane so that it was possible for them to strike each other when caused to vibrate at large amplitudes, thereby increasing the normal overload distortion. This difficulty is avoided in later types by mounting the two ribbons in parallel planes separated by about the thickness of one ribbon. In addition to preventing distortion due to ribbons striking each other, this "biplanar" construction reduces the possibility of damage to ribbons through overmodulation.

When the light-valve ribbons are focused directly on the moving film the exposure at high frequencies does not correspond to the motion of the ribbons if the velocity of either edge of the image is comparable to that of the film. The exposure given to the film is

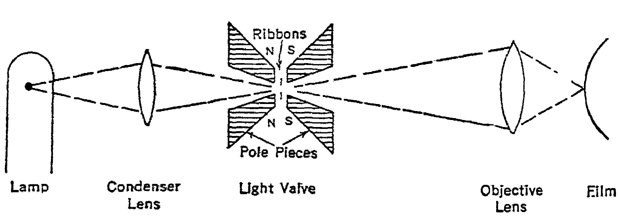


Fig. 3. Optical Schematic of Typical Light-valve Recording System

determined by the time required for any point on the film to pass through the image of the light-valve slit; in other words, the exposure (to a first approximation) is the product of time and intensity. If the frequency being recorded is low, so that the velocity of the ribbons is small compared with that of the film, the variations in film exposure will accurately correspond to the light-valve modulation. As the frequency becomes higher, however, the velocity of the ribbons increases, being substantially proportional to frequency for constant electrical power input to the light valve, until at the highest audio frequencies the ribbon-velocity effect may become important. For a sine-wave signal just sufficient to fully modulate a two-ribbon valve spaced to 0.001 in., the peak ribbon velocity becomes equal to the film velocity at about 6000 cycles. This results in a loss of high-frequency response which also varies with the average ribbon spacing. Consequently, when a high-frequency wave is superposed on a low-frequency wave, the high-frequency response varies during each cycle of the low frequency.

Distortion of this type is not indicated by harmonic measurements since it occurs mainly at frequencies whose harmonics would be greatly attenuated by other elements in the system, such as the width of the reproducing slit. Such distortion can, however, be detected by "intermodulation" tests in which two frequencies are recorded simultaneously (one high and one low) and the variation in high-frequency response measured. In usual practice the amplitude of the low frequency (60 cycles) is four times that of the high frequency (7000 cycles). The reproduced wave is passed through a high-pass filter, eliminating the 60-cycle component. When this wave is rectified a new set of low frequencies (60, 120, 180... cycles) is produced, owing to the presence of distortion frequencies in the high-frequency wave. The average amplitude of this new low-frequency wave is a convenient measure of the distortion. In a suitably designed light-valve modulator the intermodulation distortion may be reduced to 4 per cent or less, corresponding to approximately 1 per cent harmonic distortion as measured in a system in which distortion and transmission do not vary with frequency. This is accomplished by reducing the effective image width to 0.00025 in. or less, usually by means of a short-focus cylindrical lens placed close to the film surface.

NOISE-REDUCTION SYSTEM. A noise-reduction system is commonly used in film recording in order to lower the level of background noise when it is not masked by relatively loud sounds. This system is based on the fact that in a variable-density sound track the background noise output from a light print is greater than that from a dark print when reproduced with the same gain setting of the amplifier. However, when a print is made darker by merely increasing the exposure in the printer, both the ground noise and the wanted sound are reduced in approximately the same ratio so that no improvement in the signal-to-noise ratio results. The desired result is obtained by reducing the average exposure of the negative during periods of low modulation without reducing the amount of light modulated by the signal.

In the Western Electric system of noiseless recording the mean spacing of the light-valve ribbons is made to vary so that as modulation is impressed on the valve the mean spacing increases sufficiently to accommodate the increasing input. The spacing between light-valve strings is mechanically adjusted to 0.001 in., and a source of direct current is connected to the strings sufficient to reduce the spacing to 0.0003 in. (for 10-db noise reduction). In addition to this fixed bias a varying bias proportional to the envelope of the signal wave increases the ribbon spacing as the signal current increases. The lower average exposure of the negative during periods of low signal levels produces a positive with relatively high average density during such periods. The ground noise is thereby reduced when the signal is low, but, since the valve modulation due to the signal is unchanged, the effective signal-to-noise ratio is increased.

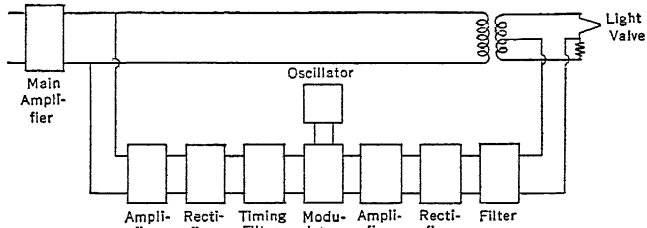


FIG. 4. Block Schematic of Noise-reduction System

A block schematic of a typical noise-reduction circuit is shown in Fig. 4. The timing filter not only serves to remove audible frequencies from the rectified signal but also shapes the bias wave so that, for rapidly fluctuating signals such as speech, the valve spacing increases at about the same rate as the signal increases, but the spacing decreases at a considerably slower rate. The oscillator, modulator, second amplifier, and rectifier merely form a convenient means of amplifying the fixed and variable bias currents.

28. REFLECTING-GALVANOMETER RECORDING SYSTEM

The reflecting-galvanometer method, commonly used for producing variable-area sound tracks, utilizes a recording device that operates on the principle of the mirror oscillograph. Light is reflected from the oscillating mirror of the recorder and is passed through a narrow slit onto the film. The resulting sound track has constant density but variable

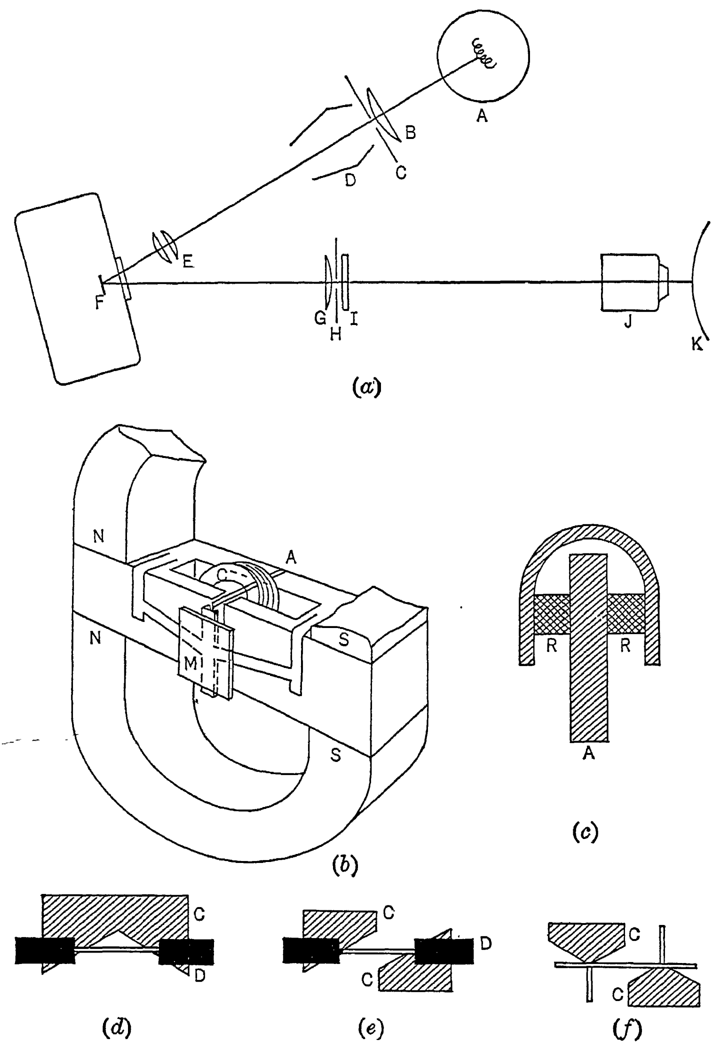


FIG. 5. (a) Optical Schematic of Typical Galvanometer Recording System. (b) Recording Galvanometer. (c) Armature Damping in Recording Galvanometer. (d) (e) (f) Typical Mask and Shutter Arrangements Used in Galvanometer Recording System.

width. A widely used type of light modulator accomplishes this by the use of a triangular beam of light which is caused to move at right angles to the axis of the slit by the recording galvanometer, so that, as it vibrates, the length of the illumined portion of the slit varies.

The RCA variable-width light modulator, Fig. 5(a), consists essentially of an incandescent lamp, a system of lenses to direct the light, an aperture and slit to limit the light, and a reflecting-mirror galvanometer to modulate the light. An image of filament *A* is formed at the galvanometer mirror *F* by the combination of lenses *B* and *E*. The aperture *C*, shown as the shaded area in Fig. 5(*d*), limits the light projected to the galvanometer mirror. Lens *E* forms an image of aperture *C* on slit *H*. Vibration of the galvanometer mirror moves this light beam back and forth across the slit so that the length of the illuminated portion of the slit is proportional to the angular deflection of the mirror. A reduced image of the slit is formed on the film *K* by objective lens *J*.

Ground noise is reduced in this modulator by cutting off the light from the ends of the slit during periods of low modulation. The two magnetically operated shutter vanes *D* are used for this purpose, producing the varying black margins shown, for example, in Fig. 1(*g*). Current for operating the noise-reduction shutters is derived from the signal and is proportional to the envelope of the signal wave. During periods of no modulation the clear area is reduced to a width of about 0.002 in.

Other common accessories in this modulator are a visual monitor system and an ultraviolet filter. The visual monitor is provided by forming an image of one edge of the aperture *C* on a monitor card. Vibration of the galvanometer mirror causes the image to lengthen in proportion to the mirror deflection. By means of suitably spaced lines on the card the operator is enabled to judge when the galvanometer deflection reaches the maximum allowable without exceeding the sound-track width (0.076 in.) that can be scanned by the slit in the reproducing machine. The ultraviolet filter *I* restricts the film exposure to a narrow band of wavelengths in the neighborhood of 3650 A, thereby reducing image spread due to scattering of light in the film emulsion. Recent fine-grain films have also contributed to improvement in image quality.

The recording galvanometer, Fig. 5(*b*), consists of a pair of permanent magnets, pole pieces, balanced armature *A*, signal and bias coils *C*, mirror support, and mirror *M*. A pair of phosphor-bronze springs hold a groove in the ribbon support against a knife edge on the end of the armature. Current through either the signal or biasing coils polarizes the armature, causing it to be attracted to the pole piece of opposite polarity and by its lateral motion rotating the mirror support and mirror.

One of the serious problems in electromechanical apparatus is the provision of suitable damping. Oil damping, although widely used in oscillograph galvanometers, has a number of objections, among which are the change of viscosity with temperature, relatively large mass required for the damping obtained, and difficulty of avoiding leakage. The damping in this galvanometer is obtained by utilizing the properties of tungsten-loaded rubber so mounted that damping is obtained in the frequency range near armature resonance, but without affecting the low-frequency response of the galvanometer. As shown in Fig. 5(*c*), two small pieces of tungsten-loaded rubber *R* are cemented to the sides of the armature *A*, and a bronze yoke presses against their outer faces. At low frequencies the yoke and pads move with the armature, but at high frequencies the inertia of the yoke causes it to tend to stand still while the armature vibrates inside it, compressing the rubber and damping the peak.

Other types of sound tracks can be made with the same modulator by means of masks of other shapes. For example a class *A* pushpull variable-width track may be made with the mask in Fig. 5(*e*). Class B pushpull variable-width tracks may be made with the mask shown in Fig. 5(*f*). Variable-density tracks can also be made by a modification of this modulator in which the uniformly illuminated triangles are replaced by a penumbra designed to give a linear gradation of light intensity. The galvanometer mirror causes this penumbra to move across the fixed slit, varying the light transmitted to the film in proportion to the mirror deflection.

29. FLASHING-LAMP AND KERR CELL RECORDING SYSTEMS

Although not widely used at the present time, the flashing-lamp method of recording has the advantage of extreme simplicity. The gaseous discharge is concentrated in a relatively small area and so designed that the current is proportional to the signal voltage. An early type of flashing lamp, called Aeolight, is a two-element tube containing an inert gas such as helium. One of the elements is of nickel; the other is coated with barium and strontium oxides. When sufficient voltage is applied to the electrodes, ionization takes place and a concentrated glow appears at the cathode. The intensity of the light so produced increases in proportion to the increase in applied voltage. In operation, sufficient polarizing voltage is applied to the tube to give the required average exposure of the negative, and sound voltages are superposed on this polarizing voltage. Since the light output is comparatively low, the lamp is usually arranged for direct illumination of the film rather than by means of the usual lens system. A fixed slit 0.0008 in. by 0.100 in., made by engraving the silvered surface of a small quartz block, is located within less than 0.001 in. of the film. The low light intensity obtained from flashing-lamp light modulators results in a very low value of negative exposure and is in the region known as the "toe" of the *H* and *D* curve. Prints made from these negatives must be printed on the "toe" of the positive *H* and *D* curve in order to minimize distortion. Consequently the permissible modulation ratio is reduced and the volume range is less than for the fully exposed type of record.

Another relatively low-intensity light modulator utilizes the Kerr electro-optical effect. A glass cell containing nitrobenzol is placed between two Nicol prisms, and an electrostatic field is applied to electrodes on either side of the cell. Light from an incandescent lamp is plane polarized by the first Nicol prism, and the second prism is so arranged that, when no polarizing voltage is applied to the cell electrodes, no light is transmitted. Since the plane of polarization of light transmitted through the cell is rotated as the polarizing voltage is increased, the transmitted light also is increased. However, the relation between transmitted light and applied voltage is not linear and introduces appreciable distortion at high modulation.

30. SOUND-ON-FILM REPRODUCING SYSTEMS

In the reproduction of either of the two normal types of single sound tracks, i.e., variable-density or variable-area, the light from an incandescent filament is made to fall on a fixed slit which is imaged on the film as a line of light 0.0012 by 0.084 in. Two common optical systems are shown in Fig. 6. In either, the light passing through the film falls directly on a photoelectric cell or is transmitted to it through a suitable combination of lenses and prisms. Usually the photocell optical system subtends a solid angle of less than 2π radians as measured from a point on the sound track. Light passing through the film emulsion is scattered through a variable solid angle depending on the density (Callier effect), so that at the higher densities a smaller proportion of the transmitted light reaches the photocell. This gives the effect of increased density gradation (gamma) and must be taken into account in processing variable-density sound tracks.

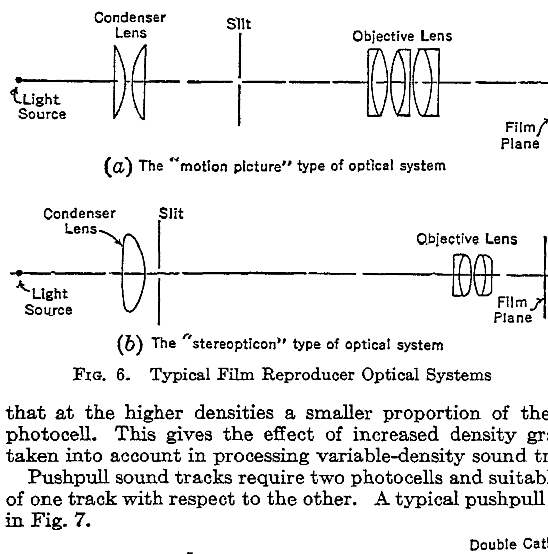


FIG. 6. Typical Film Reproducer Optical Systems

Pushpull sound tracks require two photocells and suitable means for reversing the phase of one track with respect to the other. A typical pushpull reproducer, simplified, is shown in Fig. 7.

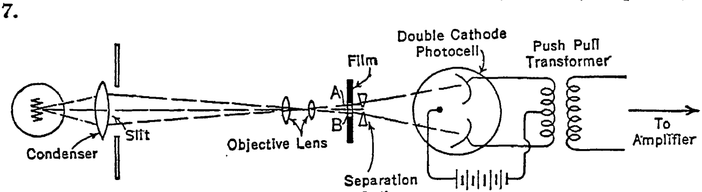


FIG. 7. Push-pull Reproducer Optical System

In the reproduction of photographic sound tracks there are unavoidable losses in output related to the speed with which the sound track is moved, and the dimensions and relative positions of recording and reproducing slits. Both the recording and the reproducing light beams should be exactly perpendicular to the direction of the film motion. Any deviation from this direction is called an error in azimuth and introduces a loss at high frequencies. Another important loss is due to the impossibility of producing on the sound track a line of light which is infinitely narrow. These conditions produce loss of efficiency at higher frequencies and in variable-area tracks may also introduce distortion. Figure 8 shows typical loss curves of these effects plotted from the following equations.

VARIABLE-DENSITY SOUND TRACKS. Effect of image width for zero azimuth:

$$\text{Scanning loss in decibels} = 20\log_{10}\left[\frac{\sin\left(\dfrac{2\pi f\beta}{V}\right)}{\dfrac{2\pi f\beta}{V}}\right] \tag{1}$$

where f = frequency in cycles per second, β = half image width, and V = film velocity.

Effect of azimuth for fixed image width:

$$\text{Scanning loss in decibels} = 20\log_{10}\left[\frac{\sin\left(\dfrac{2\pi f\beta\sec\alpha}{V}\right)}{\dfrac{2\pi f\beta\sec\alpha}{V}}\cdot\frac{\sin\left(\dfrac{2\pi fl\tan\alpha}{V}\right)}{\dfrac{2\pi fl\tan\alpha}{V}}\right] \tag{2}$$

where l = half width of sound track, and α = azimuth deviation angle.

If the azimuth is zero the last equation reduces to the equation preceding it. For very small angles (0° to 6°), which covers the cases of practical interest, $\beta\sec\alpha = l\tan\alpha$. Hence in the azimuth-loss equation the image-width factor becomes equal to the azimuth factor when the azimuth deviation is equal to the effective slit image width; i.e., the scanning loss for a particular azimuth deviation is equal to that of an image width of the same distance along the length of the film.

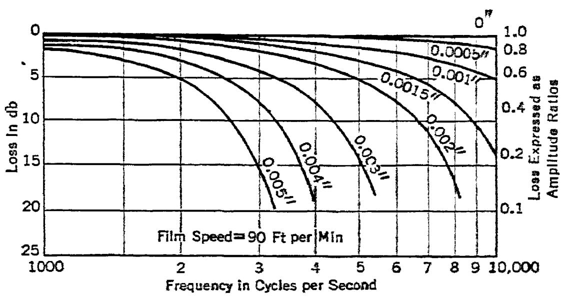


FIG. 8. Loss Due to Finite Slit Width

VARIABLE-AREA SOUND TRACKS. For variable-area sound tracks the calculation of the effect of the aperture on fundamental response and generation of harmonics becomes considerably more complicated and involves a number of assumptions that are not always realized in practice. However, the general effects may be obtained by assuming an ideal record in which the exposed portion has uniform density such as would be realized with an infinitely narrow recording slit and an ideal film emulsion. The effect of finite reproducing-slit width (with no azimuth error) is a reduction in response at high frequencies exactly the same as shown above for variable-density sound tracks. On the other hand, when the same ideal variable-area record is reproduced by means of a finite slit not perpendicular to the direction of motion, the loss in fundamental amplitude for a bilateral track is given by

$$L = \frac{1}{2am}k_1\frac{\sin\epsilon}{\epsilon} \tag{3}$$

and the ratio of second harmonic to fundamental by

$$R = \frac{k_2}{k_1}\cos\epsilon \tag{4}$$

where 2ϵ = width of reproducing image, a = width of unmodulated unbiased track, and m = per cent modulation, and

$$k_1 = 4pJ_1\left(\frac{am}{p}\right)\cdot\cos\left(\frac{a}{p}\right) \tag{4a}$$

$$k_2 = 2pJ_2\left(\frac{2am}{p}\right)\cdot\sin\left(\frac{2a}{p}\right) \tag{4b}$$

and p = cotangent of angle of azimuth deviation; also J_1 and J_2 are Bessel functions of the first and second order respectively.

Third and higher harmonics are also produced by an azimuth deviation of the reproducing slit. Distortion may also be caused by the finite width of the recording slit, but it may be compensated to some extent by choosing a suitable density for the print. The photographic image spread in the positive, then, to a first approximation, compensates for both the finite recording slit width and the negative image spread.

Distortion may also occur in the reproduction of variable-area sound records as the result of uneven illumination along the length of the reproducing slit. Although the amount of such distortion depends on the type of variable-area track and on the type and amount of unevenness of illumination, proper design of the reproducer optical system and a reasonable degree of adjustment should result in harmonics not more than 3 per cent of the fundamental (for full modulation) and in most cases considerably less.

In addition, variable-area records will obviously be distorted if the center line of the record in the reproducer is displaced from the center line of the slit image so that part of the record is not scanned. It is for this reason that 0.076 in. is considered the maximum track width for distortionless reproduction since with a standard 0.084-in. scanning image a tolerance of ±0.004 in. is then allowed for film weave.

PHOTOGRAPHIC REQUIREMENTS. The principal photographic requirement of sound recording is that the variation of light from the average amount transmitted by the film in the reproducer must be proportional to the corresponding variation from the average amount of light transmitted by the recording modulator. This general relation is true for both variable-density and variable-area tracks.

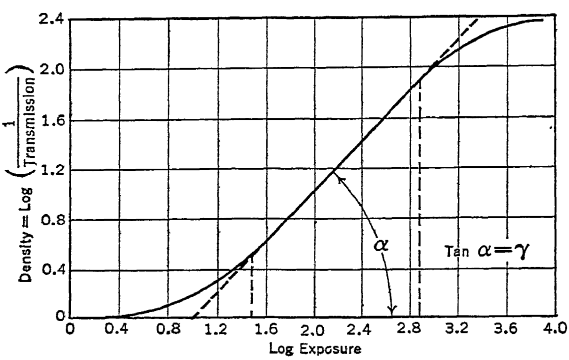


FIG. 9. *H* and *D* Curve

For a variable-density track, the exposure varies from point to point along the length of the track. Figure 9 shows the manner in which exposure and density are related in a typical film emulsion. The exact shape, and particularly the slope of the central portion of the *H* and *D* curve, depend on the development as well as the photographic properties of the film, but most *H* and *D* curves have a portion that is essentially straight when plotted as shown. The slope of the straight portion is called gamma, γ. If, when a negative is printed, a similar curve is drawn relating the exposure of the negative to the density of the positive, the slope of the straight portion of this curve is the overall gamma. It can be shown that for minimum distortion in variable-density recording the overall gamma should be unity. In this computation factors such as the Callier effect (projection factor) and the departure of actual gammas from the values measured by a sensitometer must be taken into account. Since the positive film carries both sound and picture, both images must be given the same development. The positive gamma is chosen to be suitable for the picture and is usually about 2.2. Consequently, the negative of a variable-density sound track is developed to a gamma of about 0.4 when measured on type 2B sensitometer.

Although optimum values of gamma and density for variable-density records can be obtained with a fair degree of accuracy from sensitometric measurements, the simplest and most reliable determination of these constants is made by means of intermodulation tests. Such tests measure the non-linear distortion in a print made from a negative on which two frequencies (usually 60 and 1000 cycles) are simultaneously recorded. Not only is this form of test more sensitive than a measurement of harmonics of a single frequency but also it corresponds more closely to audible distortion in commercial records.

In processing variable-area records the aim is to obtain minimum density in the clear area and maximum density in the exposed area. These are obtained by developing both negative and positive to relatively high gammas, although excessive development may result in fog in the clear area.

Variable-width records are also subject to non-linear distortion if they are not given proper exposure and development. This distortion is frequently due to spreading of the image and produces an effect similar to rectification. Optimum processing is determined by "cross-modulation" tests in which a modulated high frequency is recorded. Distortion is measured by the amount of low frequency produced by the photographic rectification of the modulated high frequency. Distortion may also be caused by non-linearity of the modulator and may be measured by harmonic or intermodulation tests.

BIBLIOGRAPHY

All references are to the *Journal of the Society of Motion Picture Engineers* unless otherwise noted.

General References

Academy of Motion Picture Arts and Sciences, *Motion Picture Sound Engineering*. Van Nostrand, New York (1938).
Academy of Motion Picture Arts and Sciences, *Recording Sound for Motion Pictures*. McGraw-Hill, New York and London (1931).
Talbot, R. H., Some Relationships between the Physical Properties and the Behavior of Motion Picture Film, Vol. XLV, 3, p. 209 (September 1945).
Kellogg, E. W., The ABC of Photographic Sound Recording, Vol. XLIV, 3, p. 151 (March 1945).
Loye, D. P., and K. F. Morgan, Sound Picture Recording and Reproducing Characteristics, Vol. XXXII, 6, p. 631 (June 1939), and Vol. XXXIII, 1, p. 107 (July 1939).
Honan, E. M., and C. R. Keith, Recent Developments in Sound Tracks, Vol. XLI, 2, p. 127 (August 1943).

Light-valve Method

MacKenzie, D., Sound Recording with the Light Valve, *Trans. Soc. Mot. Pict. Engrs.*, Vol. XII, 35, p. 730 (September 1938).
Shea, T. E., W. Herriott, and W. R. Goehner, The Principles of the Light Valve, Vol. XVIII, 6, p. 697 (June 1932).
Frayne, J. G., and H. C. Silent, Push-pull Recording with the Light Valve, Vol. XXXI, 1, p. 46 (July 1938).

Reflecting-galvanometer Method

Dimmick, G. L., Galvanometers for Variable Area Recording, Vol. XV, 4, p. 428 (October 1930).
Batsel, M. C., and E. W. Kellogg, The RCA Sound Recording System, Vol. XXVIII, 5, p. 507 (May 1937).
Dimmick, G. L., The RCA Recording System and Its Adaptation to Various Types of Sound Track, Vol. XXXIX, 3, p. 258 (September 1937).

Kerr Cell Method

Zworykin, V., L. B. Lynn, and C. R. Hanna, Kerr Cell Method of Sound Recording, *Trans. Soc. Mot. Pict. Eng.*, Vol. XII, 35, p. 748 (September 1928).

Ground-noise Reduction

Kreuzer, B., Noise Reduction with Variable Area Recording, Vol. XVI, 6, p. 671 (June 1931).
Silent, H. C., and J. G. Frayne, Western Electric Noiseless Recording, Vol. XVIII, 5, p. 551 (May 1932).
Kellogg, E. W., Ground Noise Reduction System, Vol. XXXVI, 2, p. 137 (February 1941).
Scoville, R. R., and W. L. Bell, Design and Use of Noise Reduction Systems, Vol. XXXVIII, 2, p. 125 (February 1942).

Reproducing Systems

Cook, E. D., The Aperture Effect, Vol. XIV, 6, p. 650 (June 1930).
Stryker, N. R., Scanning Losses in Reproduction, Vol. XV, 6, p. 610 (November 1930).
Cook, E. D., The Aperture Alignment Effect, Vol. XXI, 5, p. 390 (November 1933).
Foster, D., Effect of Orientation of the Scanning Image on the Quality of Sound Reproduced from Variable Width Records, Vol. XXXIII, 5, p. 502 (November 1939).
Batsel, M. C., and C. H. Cartwright, Effect of Uneven Slit Illumination on Distortion in Several Types of Variable Width Records, Vol. XXIX, 5, p. 476 (November 1937).
Carlson, F. E., Properties of Lamps and Optical Systems for Sound Reproduction, Vol. XXXIII, 1, p. 80 (July 1939).

Photographic Tone Reproduction

Hardy, A. C., The Rendering of Tone Values in the Photographic Recording of Sound, *Trans. Soc. Mot. Pict. Eng.*, Vol. XI, 31, p. 475 (September 1927).
Jones, L. A., On the Theory of Tone Reproduction, with a Graphic Method for the Solution of Problems, Vol. XVI, 5, p. 568 (May 1931).
Mees, C. E. K., Some Photographic Aspects of Sound Recording, Vol. XXIV, 4, p. 285 (April 1935).

Measurement of Distortion

Baker, J. O., and D. H. Robinson, Modulated High Frequency Recording as a Means of Determining Conditions for Optimal Processing, Vol. XXX, 1, p. 3 (January 1938).
Frayne, J. G., R. R. Scoville, Analysis and Measurement of Distortion in Variable Density Recording, Vol. XXXII, 6, p. 648 (June 1939).

Dimensional and Other Standards

Motion Picture Standards (Z22) are obtainable from the American Standards Association, 70 E. 45 St., New York, N. Y., or from the Society of Motion Picture Engineers, 342 Madison Ave., New York, N. Y.

PIEZOELECTRIC CRYSTALS

By W. P. Mason

31. DEFINITION OF EFFECTS

A piezoelectric crystal is a crystal which suffers a change in dimension or form proportional to an applied electrical potential, for small applied potentials, and conversely generates a surface charge when subject to stresses. These properties of piezoelectric crystals allow a coupling to be made between an electrical circuit driving a mechanical circuit or with the mechanical properties of the crystals themselves used to create an electric voltage as a result of mechanical motion. When a crystal is used to couple an electrical to a mechanical system it is said to be an electromechanical transducer. An example of such a device is a crystal used as the pickup unit in a phonograph, in which function it transforms the mechanical vibrations of the record into electrical vibrations which are amplified and produce sound vibrations through the loudspeaker. When the mechanical resonances of the crystal itself are used, the crystal is said to be a piezoelectric resonator or a piezoelectric oscillator.

The piezoelectric effect was discovered in 1880 by the brothers Jacques and Pierre Curie. They discovered first the "direct" effect, which is the production of charge on a crystal surface due to the effect of a mechanical force applied to the crystal surface. They also measured the "inverse" effect, which is the change in shape of the crystal due to an applied potential. Voigt (*Lehrbuch der Kristallphysik*, B. Teubner, 1910) later showed that all the linear properties of a crystal under applied stresses, potentials, and temperatures resulted in strains, electrical displacements, and increases in heat energy according to the equations

$$
\begin{aligned}
S_i &= \sum_{j=1}^{6} s_{ij}T_j + \sum_{k=1}^{3} d_{ik}E_k + \alpha_i\,\delta\Theta \quad (i, j = 1, \cdots, 6) \\
\frac{D_l}{4\pi} &= \sum_{l=1}^{3} d_{lj}T_j + \sum_{k=1}^{3} \frac{\epsilon_{kl}}{4\pi} E_k + p_l\,\delta\Theta \quad (k, l = 1, \cdots, 3) \\
\delta Q &= \sum_{j=1}^{6} \Theta\alpha_j T_j + \sum_{k=1}^{3} \Theta p_k E_k + \rho C_p\,\delta\Theta
\end{aligned}
\tag{1}
$$

where $S_i(i = 1 \text{ to } 6)$ are the six strains that can exist in a solid body, $T_j(j = 1 \text{ to } 6)$ are the six stresses in the body, $E_k(k = 1 \text{ to } 3)$ the three potential gradients (ratio of total potential divided by distance over which they are applied) that exist along the three axes, Θ is the absolute temperature in degrees Kelvin and $\delta\Theta$ the increase in temperature, $D_l(l = 1 \text{ to } 3)$ is the electric displacements along the three axes, and δQ is the increment in heat energy due to applied stresses, fields, and temperature increments. (These symbols have now been standardized by the Institute of Radio Engineers.)

The first equation says that any one of the strains, for example S_1, is in general proportional to the six stresses, the three electric fields along the three axes, and the temperature increment $\delta\Theta$. The constants s_{ij} which relate the strains to the applied stresses are the elastic moduli of compliance. Since it can be shown that $s_{ij} = s_{ji}$ there are 21 such constants for the most general crystal, a triclinic crystal. If any elements of symmetry exist in the crystal the number of independent constants is reduced. For example, ammonium dihydrogen phosphate (ADP) has six independent elastic compliances, quartz has seven, and Rochelle salt (sodium potassium tartrate) has nine. The d_{ik} constants are the piezoelectric constants which relate the strains to the applied fields. For the most general crystal there are 18 independent constants, but for more symmetrical crystals the number is reduced. ADP has two independent constants, quartz two, and Rochelle salt three. The α_i constants are the six temperature expansion coefficients which relate the six strains to the applied temperature increase $\delta\Theta$.

The second equation states that the electric displacement is proportional to the applied stresses (the constants of proportionality being again the piezoelectric constants), to the applied fields (the constants of proportionality being the dielectric constants ϵ_{kl}), and to the increase in temperature $\delta\Theta$ (the constant of proportionality being the pyroelectric constants p_l). (The equation as written is valid for the cgs system of units. For the mks system the 4π is removed from D_l and ϵ_k.) Since $1/(4\pi)$ times the normal component of the electric displacement at the surface of the crystal is equal to the surface charge σ, this equation shows the origin of the direct piezoelectric effect (while eq. [1] expresses the inverse effect).

The third equation states that the increment of heat δQ is proportional to the stresses, the fields, and the applied temperature increment. The first effect is called the stress-caloric effect, and the constant of proportionality is the absolute temperature Θ times the temperature expansion coefficients α_j. The second effect is called the electrocaloric effect, and the constant of proportionality is the absolute temperature Θ times the pyroelectric constant p_k. This last is the ratio of the electric displacement to the applied temperature $\delta\Theta$ measured at constant stress and constant field. The last term in the third equation expresses the increase in heat energy due to a temperature increase $\delta\Theta$, and the constant of proportionality is the density ρ times the specific heat at constant stress C_p. In all these equations the constants of proportionality for one variable are measured with the other two variables in the equation held constant. Thus s_{ij}, for example, could be written with two superscripts $s_{ij}^{E,\Theta}$, indicating that in determining the constants the fields E and the temperature Θ are held constant. They are therefore the constant field, isothermal elastic compliances. Similar superscripts can be written for the other terms.

For most piezoelectric applications the vibrations are so rapid that there is no time for an interchange of heat, and adiabatic conditions prevail. This can be taken account of in eq. (1) by setting $\delta Q = 0$. If we do this and eliminate $\delta\Theta$ from the remaining equations we have two equations given by

$$\begin{aligned} S_i &= \sum_{j=1}^{6} s_{ij}^E T_j + \sum_{k=1}^{3} d_{ik} E_k \\ \frac{D_l}{4\pi} &= \sum_{j=1}^{6} d_{lj} T_j + \sum_{k=1}^{3} \frac{\epsilon_{kl}^T}{4\pi} E_k \end{aligned} \tag{2}$$

All these constants are adiabatic, and they are related to the isothermal constants of eq. (1) by the relations

$$s_{ij}^E = s_{ij}^{E,\Theta} - \frac{\alpha_i^E \alpha_j^E}{\rho C_p^E}; \qquad d_{ik} = d_{ik}^\Theta - \frac{\alpha_i p_k^T \Theta}{\rho C_p^E}; \qquad \epsilon_{kl}^T = \epsilon_{kl}^{T,\Theta} - \frac{p_k^T p_l^T \Theta}{\rho C_p^E} \tag{3}$$

where the constants on the left side are understood to be adiabatic. Equations (2) represent all the linear adiabatic relations existing for a piezoelectric crystal, while eqs. (1) give all the isothermal linear relations existing for a piezoelectric crystal. Two second-order effects, the piezo-optical, and the electro-optical, are also of some interest, but they will not be discussed here. They result from a change in the dielectric constant as a function of applied stresses and applied fields respectively.

For ferroelectric crystals such as Rochelle salt the constants in the equations of the form (2) go through very wide variations over a temperature range. It has been found * that, if the electric displacement is used as the dependent variable instead of the field, the resulting constants are nearly independent of temperature. These relations can be obtained from eqs. (2) by solving in terms of D_l and T_j; they are

$$\begin{aligned} S_i &= \sum_{j=1}^{6} s_{ij}^D T_j + \sum_{l=1}^{3} g_{il} \frac{D_l}{4\pi} \\ E_k &= -\sum_{j=1}^{6} g_{kj} T_j + \sum_{l=1}^{3} \beta_{kl}^T D_l \end{aligned} \tag{4}$$

where

$$\beta_{kl}^T = \frac{(-1)^{k+l}\Delta^{k,l}}{\Delta}; \qquad g_{kj} = \beta_{k1} d_{1j} + \beta_{k2} d_{2j} + \beta_{k3} d_{3j}$$

$$s_{ij}^D = s_{ij}^E - [d_{i1}g_{j1} + d_{i2}g_{j2} + d_{i3}g_{j3} + d_{i4}g_{j4} + d_{i5}g_{j5} + d_{i6}g_{j6}]$$

and Δ is the determinant

$$\Delta = \begin{vmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{12} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{13} & \epsilon_{23} & \epsilon_{33} \end{vmatrix}$$

and $\Delta^{k,l}$ is the determinant obtained by suppressing the kth row and lth column.

A ferroelectric crystal is one which shows a spontaneous polarization over a given temperature range. This is due to the movable electric dipoles exerting a mutual reaction and lining up for one direction of the crystal. The effect is similar to the ferromagnetic effect in magnetic substances and is accompanied by similar effects. The polarization vs. potential curves show hysteresis effects and very high dielectric constants. Large piezoelectric effects exist in the ferroelectric range, and Rochelle salt, for example, has a d_{14} piezoelectric constant which may be 1000 times as large as that for a quartz crystal. The limiting temperatures for the ferroelectric regions are called the Curie temperatures. For Rochelle salt these are -18 and $+24$ deg cent. Other ferroelectric crystals are also known, notably potassium dihydrogen phosphate and potassium dihydrogen arsenate (see Busch, "Neue Seignette Elektrica," *Helv. Phys. Acta*, II No. 3 [1938]), but their Curie temperatures are very low, namely, -151 and -182 deg cent. Rochelle salt was the first crystal discovered that had its ferroelectric region in the room-temperature range. This fact accounts for its wide use in acoustic devices in spite of its poor mechanical and

* See W. P. Mason, A Dynamic Measurement of the Elastic, Electric, and Piezoelectric Constants of Rochelle Salt, *Phys. Rev.*, Vol. 55, 775 (1939), and H. Mueller, Properties of Rochelle Salt, *Phys. Rev.*, 57, 829 (1940).

chemical properties. Another ferroelectric crystal, barium titanate, has recently been discovered which has a large electrostrictive effect. This crystal in ceramic form may be an important electromechanical transducing element (Mason, "Electrostrictive Effect in Barium Titanate Ceramics," *Phys. Rev.*, Vol. 74, No. 9, pp. 1134–1148, Nov. 1, 1948).

32. APPLICATION OF PIEZOELECTRIC CRYSTALS

The piezoelectric effect remained a scientific curiosity from the time of its discovery until the time of World War I, 1914–1918. During that time Professor Langevin in Paris devised an underwater sound-locating device using quartz crystals to convert alternating electrical energy into sound vibrations in water. The sound beam was sent out into the water and was reflected back from an object or the bottom of the ocean. This reflection impinged on the crystal transducer and generated an electrical voltage which could be detected by vacuum-tube devices. This use was a forerunner of the fathometers and underwater sound-locating devices that have been widely used by the Navy. Although quartz was originally used for this purpose it has been displaced by Rochelle salt and particularly a new crystal developed during World War II, the ammonium dihydrogen phosphate or ADP crystal. This crystal has so many mechanical, chemical, power-handling capacity, and temperature advantages over Rochelle salt that it appears likely to replace all other transducing elements for underwater sound applications.

In 1922 it was shown by Professor Cady of Wesleyan University that very stable oscillators could be obtained by using quartz crystals as the frequency-controlling element. These have been applied to controlling the frequency of broadcasting stations and radio transmitters in general. Quartz crystals using some one of the low-temperature-coefficient crystals described in article 33 produce the most stable oscillators and the best time-keeping systems that can be obtained. The use of crystals to stabilize oscillators was so prevalent during World War II that over 30,000,000 crystals were produced in a single year for this purpose.

Another application of piezoelectric crystals is in producing very selective filters. On account of the very high Q existing in crystals they can practically eliminate the effect of dissipation in filter structures. Such filters have been widely applied in the long-distance telephone lines and in single-sideband transatlantic radio telephone systems. Narrow-band crystal filters have been used in picking off single frequencies and narrow bands of frequencies for control and analyzing purposes. For this application quartz crystals have been mostly used. However, it appears that the requirements are lenient enough to allow some of the synthetic crystals to be employed.

Besides producing and detecting sound in liquids, crystals have been used to produce and detect vibrations in gases and solids. On account of their high mechanical and electrical impedances crystals are at somewhat of a disadvantage in coupling to low-mechanical-impedance air waves. By using bimorph types of units which employ bending or flexural vibrations the mechanical impedances of crystals can be lowered. For sound pickup devices the high electrical impedance is not a disadvantage, for they can be worked directly into the grid of a vacuum tube which inherently is a high impedance. Hence large numbers of crystals have found uses in microphones. For this purpose Rochelle salt is common, but the constants of ADP are favorable enough so that they may displace it.

Crystals have also been used in receivers, relays, oscillographs, and other devices for which displacements are required for a given applied voltage. For this purpose, crystals having large d piezoelectric constants are required, and Rochelle salt is universally used. On account of the large variation of d with temperature such devices are not very stable and reproducible and hence are unsuitable for high-quality equipment.

Crystals have also been employed in producing very high-frequency vibrations in gases, liquids, and solids. For this purpose quartz is the almost universal choice since it can be ground very thin and can be used to produce high frequencies. X-cut quartz is utilized to set up longitudinal vibrations and Y-cut quartz to produce shear vibrations. Such high-frequency sound waves are applicable for testing steel castings and other solid materials for flaws (Firestone, The Supersonic Reflectoscope, *J. A.S.A.*, Vol. 17, No. 3 [January 1946]). They have also been utilized to study the properties of liquids, gases, and solids and the way in which they vary with frequency.

33. PROPERTIES OF QUARTZ

There are at least 500 crystalline substances that have been tested and found to be piezoelectric, and it is to be presumed that among the many thousands of compounds that will form into crystals in the 20 out of the 32 crystallographic classes that may be piezoelectric most of them will show some piezoelectric activity. However, only three crystals have received wide application in practical devices: quartz, Rochelle salt, and ammonium dihydrogen phosphate (ADP). A fourth crystal, tourmaline, has received a limited application in sound-measuring devices because it is sensitive to hydrostatic pressures. It is to be expected, however, that, with several large laboratories actively engaged in investigating new piezoelectric materials, many more crystals will eventually find practical application.

It is the purpose of the following sections to discuss the properties and useful cuts of quartz, Rochelle salt, and ADP.

PHYSICAL PROPERTIES OF QUARTZ. Quartz is described by the chemist as silicon dioxide, SiO_2, and it crystallizes in the trigonal trapezohedral class. The Z or optic axis is an axis of threefold symmetry; i.e., if one measures any property of the crystal at a definite position in the crystal, this property will be repeated at angles of $\pm 120°$ rotation about the Z axis. The melting point of quartz is 1750 deg cent, the density 2.65, and the hardness is 7 on Mohs' scale. Under atmospheric pressure, α or low-temperature quartz transforms into β or high-temperature quartz at 573 deg cent. Under stress this transformation temperature is lowered. Alpha quartz is insoluble in ordinary acids but is decomposed in hydrofluoric acid and in hot alkalies. Quartz is soluble to some extent in water at high pressures and temperatures. In an enclosed system, crystalline quartz will dissolve in water to the extent of 3 grams per liter at 350 deg cent. Powdered fused quartz, which has a larger surface-to-volume ratio, will dissolve to a considerably larger extent.

Quartz is found principally in Brazil in several different types of deposits (see Stoiber, Tolman, and Butler, Geology of Quartz Crystal Deposits, *Am. Mineralogist*, Vol. 30, 245–268 [1945]). The preponderance of the crystals found is in the lower-weight class as shown by the table. Most of the clear quartz has recognizable natural faces, but some, particularly river quartz, has no natural faces.

Crystal Weight Groups weight in grams	Percentage of the Total Number of Crystals Which Were in Each Weight Group
200– 300	55.5
300– 500	29.5
500– 700	10.4
700– 1,000	2.1
1,000– 2,000	1.8
2,000– 3,000	0.5
3,000– 4,000	0.2
4,000– 5,000	<0.1
5,000– 7,000	<0.1
7,000–10,000	<0.1

Quartz occurs in optical right-hand and left-hand forms; i.e., the crystal will rotate the plane of polarization of polarized light passing along the Z or optic axis counterclockwise (left handed) or clockwise (right handed) from the point of view of the observer facing the source of light. Most crystals have sections with both handedness. In general, the middle section is likely to be all of one hand while the outside sections may have parts of each handedness. A conoscope may be used to locate the optic axis and will also show the handedness and position of any optical twinning. The principle of the conoscope is shown by Fig. 1. Light from the source is sent through a polarizer and through the converging lens L_1. This lens sends converging or conical beams through the crystal which are gathered by the second lens, focused, and sent through the analyzer. In practice the lenses and crystal are immersed in a liquid having the same index of refraction as the crystal along its optic axis. Such liquids may be mixtures of Decalin and Dowtherm or dimethyl phthalate and α monochlor naphthalene. The crystal breaks up all rays not parallel to the optic axes into two components which travel with different velocities. Hence the analyzer is not able to extinguish the light that has traversed the crystal except at angles for which the two rays are in opposite phase. Hence one sees a series of rings in the conoscope when the direction of the Z axis is along the line between the source and the eye. Owing to the rotation of the plane of polarization in the crystal one finds that the rings either expand or contract for a right- or left-handed crystal respectively for a clockwise rotation of the analyzer. This gives a method of determining the handedness of the crystal. Optical twinning also shows up in a viewing system of this type, for it deforms the ring pattern. If plane rays rather than conical rays are used, and a source of white light, color effects also show up the position of the optical twinning.

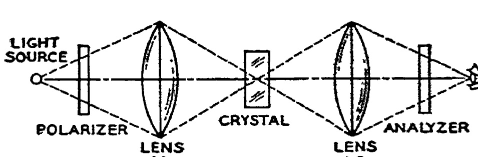


FIG. 1. Principle of Conoscope

The Dauphine or electrical type of twinning also exists in quartz. It results from a 180° change in the direction of the crystal atomic arrangements. As shown by Fig. 2, the silicon atoms normal to the Z axis are arranged in near hexagons all pointing in one direction. If the temperature is raised above 573 deg cent a change in the arrangement to the hexagonal pattern shown in the middle occurs. The result is high-temperature or β quartz. As the temperature is decreased below 573 deg, the crystal may return to the form at the bottom, or part of it may return to this form and part to the form in which the near hexagons point in the opposite direction. If both forms exist the crystal is said to have electrical twinning. The best method of detecting electrical twinning is by etching the surface with hydrofluoric acid, which eats away the crystal at rates depending on the orientation of the crystal surface. Since the two twinned areas will develop etch pits that point in opposite directions, grazing light will cause one part to reflect brightly while the other reflects diffusely, and hence one can see the parts that have different regions of electrical twinning. Since the piezoelectric effect is opposite for the two twinned areas, it is necessary that there be only one region in a useful crystal. Electrical twinning usually occurs in an untwinned plate if it is taken above the inversion point. It may also occur at lower temperatures if stress is applied. Such twinning has been observed when a hot soldering iron is pressed against a crystal, or it may even occur when the crystal is sawed. Wooster (*Nature*, Vol. 157, No. 3987 [March 30, 1946]) has found that the electrical twinning can be removed by exerting a twist around the Z or Z' axis and heating the crystal nearly to the inversion point.

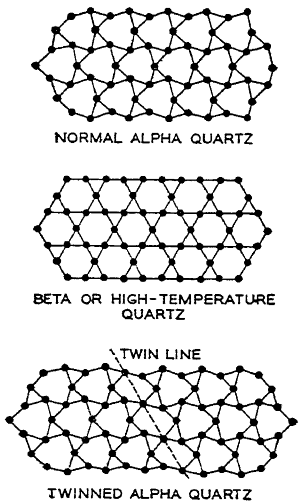


FIG. 2. Arrangement of Silicon Atoms in Quartz Normal to Z Axis

Other defects in quartz crystals are (1) "bubbles": bubble-like cavities which may be fine or large; (2) veils, heavy or fine, which are more or less continuous sheets of small bubble-like cavities; (3) clouds or haze: aggregates of fine bubble-like cavities; (4) ghosts or phantoms: outlines of earlier growths within the crystal, usually marking what were once edges of adjoining faces which become visible when a beam of light is reflected from the minute fractures or parting planes that outline them; (5) fractures. All these defects can be observed by shining a strong light through the crystal at right angles to the direction of observation. The crystal is usually immersed in an inspection tank which is filled with a liquid having the same index of refraction as the crystal. Opinions differ on how many inclusions or bubbles of a small size can be tolerated in the finished crystal.

All the inspection and orienting instruments as well as the methods of sawing and preparing the crystals are completely described in R. A. Heising, *Quartz Crystals for Electrical Circuits*, Van Nostrand, 1946, and in the May–June 1945 issue of the *American Mineralogist.*

USEFUL CRYSTAL ORIENTATIONS. The modes of motion and the properties of these modes depend markedly on how crystal plates are oriented with respect to the natural crystal faces. Figure 3 shows a natural quartz crystal, the three crystallographic axes, and some of the more important special cuts that have found use in the radio and telephone art. The Z or optic axis of the crystal is along the long direction of the crystal; the X axis lies through one of the apexes of the hexagon; and the Y axis is normal to the other two in a right-handed system. The piezoelectric, elastic, and dielectric equations of quartz take the form

$$
\begin{aligned}
S_1 &= s_{11}{}^E T_1 + s_{12}{}^E T_2 + s_{13} T_3 + s_{14}{}^E T_4 + d_{11} E_x \\
S_2 &= s_{12}{}^E T_1 + s_{11}{}^E T_2 + s_{13} T_3 + s_{14}{}^E T_4 + d_{11} E_x \\
S_3 &= s_{13} T_1 + s_{13} T_2 + s_{33} T_3 \\
S_4 &= s_{14}{}^E T_1 - s_{14}{}^E T_2 + s_{44}{}^E T_4 + d_{14} E_x \\
S_5 &= s_{44}{}^E T_5 + 2 s_{14}{}^E T_6 - d_{14} E_y \\
S_6 &= 2 s_{14}{}^E T_5 + 2(s_{11}{}^E - s_{12}{}^E) T_6 - 2 d_{11} E_y \\
\sigma_x &= \frac{D_x}{4\pi} = d_{11} T_1 - d_{11} T_2 + d_{14} T_4 + \frac{\epsilon_1{}^T}{4\pi} E_x \\
\sigma_y &= \frac{D_y}{4\pi} = -d_{14} T_5 - 2 d_{11} T_6 + \frac{\epsilon_1{}^T}{4\pi} E_y \\
\sigma_z &= \frac{D_z}{4\pi} = \frac{\epsilon_3}{4\pi} E_z
\end{aligned}
\qquad (5)
$$

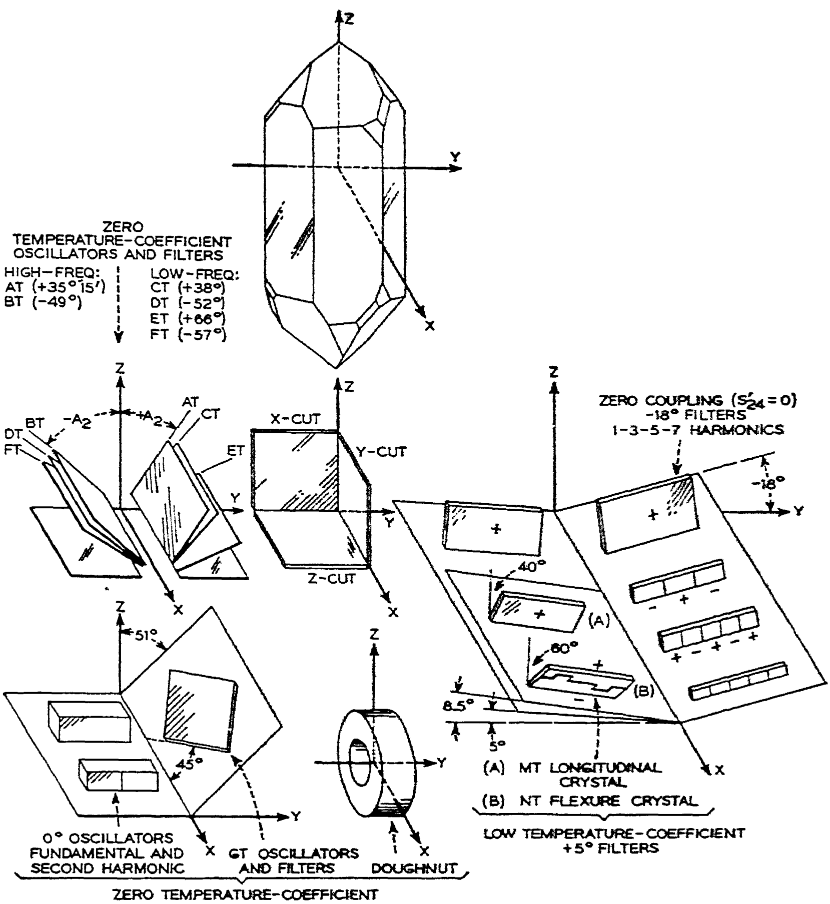


FIG. 3. Principal Cuts of Quartz

where S_1, S_2, S_3 are the three elongation strains along the X, Y, and Z axes respectively; S_4, S_5, and S_6, the three shearing strains; T_1, T_2, T_3, the three tensional stresses; T_4, T_5, and T_6, the three shearing stresses; E_x, E_y, E_z, the three fields; D_x, D_y, D_z, the three electrical displacements which at the outer surfaces are equal to the surface charge $4\pi\sigma_x$, $4\pi\sigma_y$, and $4\pi\sigma_z$. In cgs units the elastic, piezoelectric, and dielectric constants have the values (see Mason, "Quartz Crystal Applications," *B.S.T.J.*, Vol. XXII, No. 2 [July 1943]):

$$
\begin{aligned}
s_{11}^E &= 127.9 \times 10^{-14} \times \text{cm}^2/\text{dyne} \\
s_{12}^E &= -15.35 \\
s_{13}^E &= -11.0 \\
s_{14}^E &= -44.6 \\
s_{33} &= 95.6 \\
s_{44}^E &= 197.8 \\
s_{66}^E &= 2(s_{11}^E - s_{12}^E) = 286.5 \times 10^{-14}
\end{aligned}
\qquad (6)
$$

$$d_{11} = -6.76 \times 10^{-8}\,\text{statcoulomb/dyne} \qquad \epsilon_1^T = 4.58\,\frac{4\pi\ \text{statcoulomb}}{\text{statvolt}}$$

$$d_{14} = 2.56 \times 10^{-8} \qquad \epsilon_3 = 4.70$$

for the mks system the elastic compliances are multiplied by 10, the piezoelectric constants divided by 30,000, and the dielectric constants multiplied by the factor 8.85×10^{-12}.

X-CUT CRYSTALS. These equations are useful in predicting the type of motion that will be generated in a given type of cut and the magnitude of the electromechanical coupling. For example, the first equation of eq. (5) shows that a strain S_1, which is an elongation along the X axis, will be generated by a field applied along the X axis. The applied field will then generate a thickness longitudinal mode since the motion is in the same direction as the applied field. If the thickness is made small this type of crystal can produce a very high frequency, and it was originally used to control oscillators. Because of their poor temperature coefficient, such crystals have largely been replaced in the control of oscillators by *AT* and *BT* thickness shear mode crystals which have much better properties. *X*-cut crystals, however, will produce ultrasonic vibrations in solids, liquids, and gases, and such waves have been used in studying the properties of these materials and also in flaw detectors (see Firestone, *J. A.S.A.* [January 1946]) which determine whether any cracks or irregularities occur in metal castings. For this purpose it is desirable to transform as much input electrical energy as possible into mechanical energy. A measure of the efficiency of this conversion for statically or slowly varying applied fields is the electromechanical coupling factor k, which is defined by the equation

$$k = d_{11}\sqrt{\frac{4\pi c_{11}{}^E}{E_1}} = 0.095 \tag{7}$$

where $c_{11}{}^E$ is the effective elastic constant for a thickness mode. This is equal to

$$c_{11}{}^E = 8.60 \times 10^{11} \quad \text{dynes per cm}^2 \tag{8}$$

Inserting the values given in eq. (7), we find that the coupling is about 9.5 per cent. This means that, for a static field, the square of k or about 1 per cent of the input energy is stored in mechanical form. For alternating fields near the resonance of the crystal a considerably larger part, in fact, nearly all, can be converted into mechanical energy if the shunt capacity is tuned by a coil, but, nevertheless, the coupling is a measure of the width of the frequency range for which this conversion can be done efficiently. If f_B is the highest frequency and f_A the lowest frequency for which the loss is not more than 50 per cent it can be shown that

$$\frac{f_B}{f_A} = \sqrt{\frac{1+k}{1-k}} \tag{9}$$

Some synthetic crystals such as lithium sulfate and *L*-cut Rochelle salt have coupling factors of 0.35 to 0.4 and are to be preferred when it is desired to radiate a wide band of frequencies, but for high frequencies *X*-cut quartz is commonly used on account of its excellent mechanical properties.

The second equation of (5) shows that a strain S_2, which is an elongation along the Y axis, is excited when a field is applied along the X axis. Since the long direction of the crystal is taken along this direction this mode of motion is called a length longitudinal mode. It has been used to some extent to drive low-frequency oscillations in gases, liquids, and solids. Two modifications of this cut have received considerable use in the construction of quartz crystal filters. These cuts are the $-18°$ *X*-cut crystal and the $+5°$ *X*-cut crystal shown by Fig. 3. The $-18°$ cut is used because it produces a very pure frequency spectrum giving only a single resonance over a frequency range of 3 to 1 (see W. P. Mason, Electrical Wave Filters Employing Quartz Crystals as Elements, *B.S.T.J.*, Vol. XIII [July 1934]). The $+5°$ *X*-cut crystal is used because it is the best orientation of the X cuts for giving a low temperature coefficient of frequency. By putting a divided plating on the crystal as shown by Fig. 4 this crystal can be driven in a flexure mode at much lower frequencies than can be realized with a longitudinal mode. It has been used for picking off single-frequency pilot channels for controlling the gain of a carrier system.

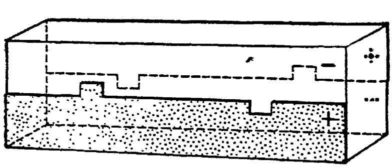


FIG. 4. Plating Arrangement for Driving a Longitudinal Crystal in Flexure

The temperature coefficient of the $+5°$ *X*-cut used for both longitudinal and flexure modes can be improved by rotating the thickness around the length of the crystal. This results in the *MT* and *NT* crystals shown by Fig. 3. These have temperature coefficients under one part in a million per degree centigrade but a smaller coupling than the equivalent $+5°$ *X*-cut crystals. (See Mason and Sykes, Low Frequency Quartz Crystal Cuts Having Low Temperature Coefficients, *Proc. I.R.E.*, Vol. 32, No. 4 [April 1944].)

Y-CUT CRYSTALS. When a field E_y is applied along the Y axis, eq. (5) shows that two types of strain are generated, S_5 and S_6. Both these strains are shearing strains which distort a square in the crystal into a rhombus as shown by Fig. 5. The S_5 strain, shown in Fig. 5, distorts the crystal in the XZ plane; the S_6 strain distorts the crystal in the XY plane. Since the field is applied along the thickness, which is the Y direction, the first strain S_5 is called a face shear strain and S_6 a thickness shear strain. The frequency of a face shear mode is controlled by the contour dimensions and hence will be relatively low. The frequency of the thickness shear mode is controlled by the thickness dimension which can be made very small and hence will result in a high frequency.

The Y-cut crystal was first used in the control of high-frequency oscillators but on account of its high temperature coefficient has largely been displaced by the AT and BT crystals which are modified Y-cut crystals. The Y-cut crystal is still used to generate shear vibrations in solids. For this purpose it has a higher coupling than the X cut, since the coupling for the shear thickness mode is

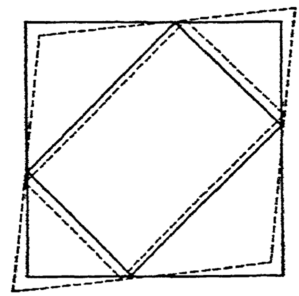


Fig. 5. Method for Obtaining a Longitudinal Vibration from a Shear Crystal

$$k = 2d_{11}\sqrt{\frac{4\pi c_{66}^E}{K_1^T}} = 0.142 \qquad (10)$$

Rotations of the thickness direction around the X axis have resulted in rotated Y cuts that have very favorable properties. Investigations made by Lack, Willard and Fair, Koga, Bechmann, and Straubel have shown how the properties of the thickness shear mode varied with angle of cut. As shown by Fig. 3 all the orientations resulting in useful crystals have their length along the X axis and their thickness makes an angle θ with the Y axis. Figure 6 shows the frequency constant (kilocycles for a crystal 1 mm thick) as a function of the angle of rotation. At an angle of rotation of $+31°$ and $-59°$ the frequency is minimum and maximum respectively. At these two angles, the mechanical coupling between the thickness shear mode, and the face shear mode and overtones, becomes zero and a crystal is obtained which is much freer from extraneous modes of motion than is the Y cut. Figure 7 shows a plot of temperature coefficient against the orientation angle, and at 35° 15′ and −49° crystals are obtained which have zero temperature coefficients. These cuts, known as the AT and BT crystals respectively, have been very widely used to control high-frequency oscillators. Frequencies as high as 10 megacycles are used for fundamental control, and by means of mechanical harmonics frequencies as high as 197 megacycles have been

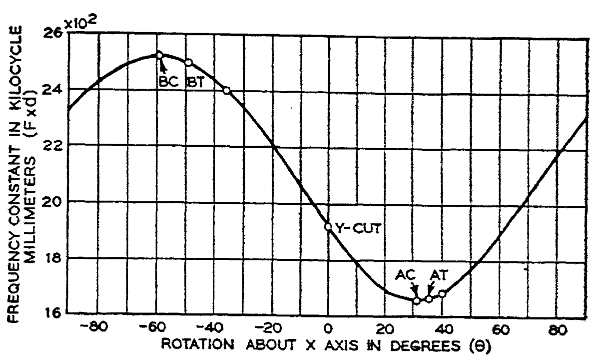


Fig. 6. Frequency Constants for Rotated Y-cut Quartz Crystals

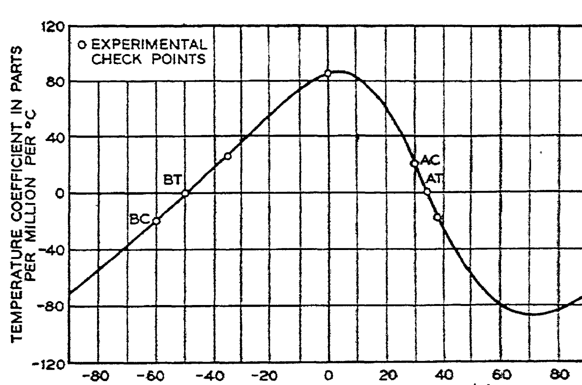


Fig. 7. Temperature Coefficients for Rotated Y-cut Quartz Crystals

obtained. (See Mason and Fair, A New Direct Crystal Controlled Oscillator, *Proc. I.R.E.*, Vol. 30, 464–472 [October 1942].)

Since the *AT* and *BT* are relatively near in angle to the *AC* and *BC* cuts they have a good frequency spectrum. Strong couplings still exist to flexure modes of motion. By measuring the modes of motion as a function of the length, width, and thickness, dimensional ratios can be obtained for which only the main mode exists for a large frequency range on either side of the main frequency. (See Sykes, Modes of Motion in Quartz Crystals, *B.S.T.J.*, Vol. XXIII, No. 1 [January 1944].) By maintaining this ratio fixed as the thickness is changed, a good crystal free from resonances over a wide temperature range is obtained. Crystals produced by the process of grinding to a set of predetermined dimensions are called predimensioned crystals and usually result in a higher-activity crystal and one having a smooth temperature frequency curve over a wide temperature range.

Another manufacturing process called the edge grinding process is sometimes employed. This consists in controlling the thickness dimension only and in removing troublesome couplings by grinding the edges of the crystal until the crystal has a high activity and is free from frequency hops over a temperature range. This process may be quicker for crystals that do not have to satisfy stringent activity and temperature requirements but is not likely to produce as satisfactory crystals as the predimensioning process. Thickness vibrating crystals may either be ground or etched to frequency. On account of an aging which appears to be due to loosely bound and misoriented layers of quartz on the surface caused by sawing and lapping processes, it has become customary to etch crystal surfaces to frequency, since this process removes the loosely bound material and leaves a surface that does not age appreciably. The aging appears to be caused by the attack of water vapor on the strained surface which results in either loosening or removing the strained material. The first process causes a lowering of the *Q* of the crystal (ratio of reactance to resistance) and a consequent lowering of the activity of the oscillator controlled by the crystal; the second process causes an increase in the frequency of the crystal. Aging can be prevented by etching the crystal surface to a depth of several microns or by hermetically sealing the crystal.

Two other methods of adjusting the frequency of crystals have been employed. One (see Sykes, High Frequency Plated Quartz Crystal Units for Control of Communications Equipment, *Proc. I.R.E.*, Vol. 34, No. 2 [February 1946]) etches the crystal frequency above the desired frequency by a predetermined number of kilocycles and then lowers the frequency by plating; by an evaporation process an amount of metal necessary to load the crystal down to its desired frequency is added to the crystal. By this method the frequency can be very accurately controlled in the final mounting. The other method utilizes the recently discovered fact that exposure to X-ray irradiation lowers the elastic constant of *BT* and *AT* crystals and hence lowers their frequency of oscillation (see Frondel, Effect of Radiation on the Elasticity of Quartz, *Am. Mineralogist*, Vol. 30 [May 1945]). The effect is produced by electrons being expelled from orbits around silicon atoms in the quartz and causing a lower energy of binding between molecules and hence a slightly lower elastic constant. This effect amounts to 0.1 per cent frequency change at the most and varies by considerable factors from crystal to crystal, presumably owing to the amount of their impurity content. Exposure to X-rays causes a darkening of the crystal, and the amount of darkening appears to be correlated with the amount of frequency change. On account of the variability of the effect, this process has not had a wide use.

Two other rotated *Y*-cut crystals that can be given zero temperature coefficients are the *CT* and *DT* face shear cuts (see Willard and Hight, *Proc. I.R.E.*, Vol. 25, 549–563 [1937]). These are nearly at right angles to the *AT* and *BT* cuts and use the same shearing moduli in the face shear mode that the *AT* and *BT* do in the thickness shear mode. The *CT* cut at $+38°$ orientation as shown by Fig. 3 has a frequency constant of 308 kc-cm for a square crystal and has been used in frequency-modulated oscillators in the frequency range from 300 to 1000 kc. The *DT* crystal is smaller for the same frequency and is used in the frequency range from 200 to 500 kc. Both these crystals had wide application in frequency-modulated oscillators for tank and artillery radio circuits during World War II.

The final rotated *Y*-cut crystal that has been used considerably for controlling very precise oscillators for time standards and in the Loran navigation system is the *GT* crystal (see Mason, A New Quartz Crystal Plate, Designated the GT, *Proc. I.R.E.*, Vol. 28, 20–223 [May 1940]). This crystal is produced, as shown by Fig. 3, by rotating the plane by $51° 7.5'$ from *Y* and by rotating the length 45° from the *X* axis. Whereas most other zero-temperature-coefficient crystals have a parabolic variation of frequency with temperature about the zero temperature coefficient as shown by Fig. 8, this parabolic variation is absent for the *GT* and a very constant frequency is produced over a wide temperature range. Hence a very moderate temperature control produces a very constant frequency. A crystal mounted by means of several wires soldered to its surface (see Greenidge, Mounting and Fabrication of Plated Quartz Crystal Units, *B.S.T.J.*, Vol. 23, 234 [July 1944])

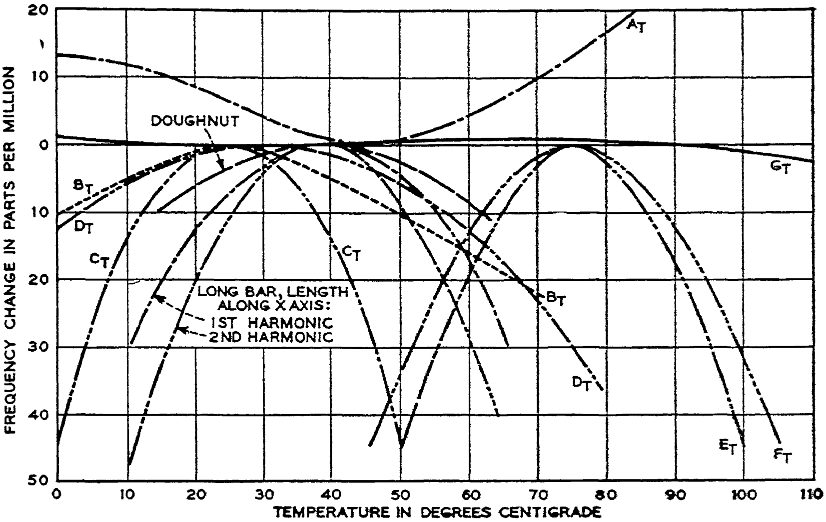


FIG. 8. Temperature Frequency Characteristics for Low-coefficient Quartz Crystals

is very stable, is little affected by shocks, and ages very little over a long period of time. It constitutes an oscillator that maintains its frequency to 1 part in 10^9 or better over long periods of time and has made possible the precise timing necessary in the Loran system and in very stable time standards (see Spencer Jones, *Endeavor*, Vol. 4, No. 16 [October 1945]).

34. PROPERTIES OF ROCHELLE SALT

Rochelle salt is sodium potassium tartrate with four molecules of water of crystallization ($NaKC_4H_4O_6 \cdot 4H_2O$) and forms in the orthorhombic bisphenoidal class. The usual form of the crystal is indicated by Fig. 9(*a*), which shows the directions of the *X*, *Y*, and *Z* axes. Since the crystal has water of crystallization it has a vapor pressure. As shown by Fig. 10, lower line, if the humidity of the surrounding atmosphere is below 35 per cent at 25 deg cent, the water-vapor pressure of the crystal is greater than the vapor pressure of water in the surrounding atmosphere and the crystal will lose water and dehydrate. This causes a white powder of dehydrated material to form on the outside of the crystal which will ruin the operation of the crystal if it becomes too large. The crystal is stable between 35 and 85 per cent relative humidity. Above 85 per cent humidity the crystal will absorb water from the atmosphere on its surface and will slowly dissolve if kept

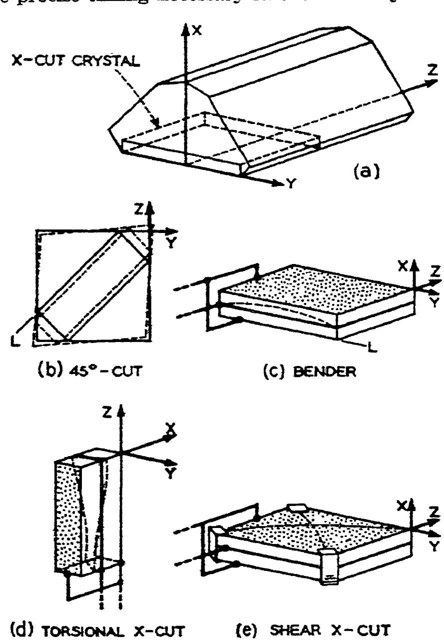


FIG. 9. Methods for Obtaining Longitudinal, Flexural, Torsional, and Plate Shear Vibrations from an X-cut Rochelle Salt Crystal

in such an atmosphere. To minimize these humidity effects the crystals are often coated with waxes, which, however, retard rather than prevent the dehydration of the crystal. If the crystal can be hermetically sealed in a container with powdered crystalline Rochelle salt and dehydrated Rochelle salt, it can be made to last indefinitely. The powdered salt will give up water if the temperature rises and the dehydrated salt will take up water if the temperature lowers, and the two will maintain a humidity that approximates the lower curve as a function of temperature. At a temperature of 55 deg cent (130 deg fahr) the crystal breaks up into sodium tartrate and potassium tartrate with the evolution of one mole of water which dissolves the two crystals in a liquid solution. If this solution is rapidly supercooled it remains quite fluid for a number of minutes before it crystallizes and hardens. This "melted" Rochelle salt forms a very stiff glue that has been used to glue together pieces of Rochelle salt.

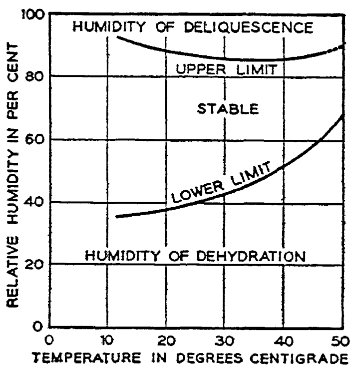


FIG. 10. Limits of Humidity Stability of a Rochelle Salt Crystal

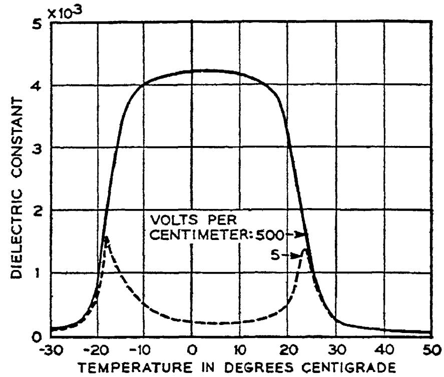


FIG. 11. Free Dielectric Constant of an X-cut Rochelle Salt Crystal as a Function of Temperature and Field Strength

Between the temperatures of -18 and $+24$ deg cent Rochelle salt has ferroelectric properties. By this is meant that it becomes spontaneously polarized in the $\pm x$ direction. A small applied field causes a large change in polarization and one which follows a hysteresis loop as does a ferromagnetic material. Since the piezoelectric strain is proportional to the polarization, a large distortion of the crystal occurs. Hence Rochelle salt is principally used when a large motion is required for a small applied voltage. The displacement, however, shows a hysteresis effect and varies considerably with temperature for a given applied voltage. The piezoelectric equations for Rochelle salt can be written in the form

$$\begin{aligned}
S_1 &= s_{11}T_1 + s_{12}T_2 + s_{13}T_3 & E_1 &= -g_{14}T_4 + \beta_{11}{}^T D_1\\
S_2 &= s_{12}T_1 + s_{22}T_2 + s_{23}T_3 & E_2 &= -g_{25}T_5 + \beta_{22}{}^T D_2\\
S_3 &= s_{13}T_1 + s_{23}T_2 + s_{33}T_3 & E_3 &= -g_{36}T_6 + \beta_{33}{}^T D_3\\
S_4 &= s_{44}{}^D T_4 + g_{14}\frac{D_x}{4\pi}\\
S_5 &= s_{55}{}^D T_5 + g_{25}\frac{D_y}{4\pi}\\
S_6 &= s_{66}{}^D T_6 + g_{36}\frac{D_z}{4\pi}
\end{aligned} \tag{11}$$

where in cgs units the constants have the values

$s_{11} = 5.18 \times 10^{-12}$ cm² per dyne
$s_{22} = 3.49 \times 10^{-12}$
$s_{33} = 3.34 \times 10^{-12}$
$s_{12} = -1.53 \times 10^{-12}$
$s_{13} = -2.11 \times 10^{-12}$
$s_{23} = -1.03 \times 10^{-12}$

$s_{44}{}^D = 7.98 \times 10^{-12}$
$s_{55}{}^D = 32.8 \times 10^{-12}$
$s_{66}{}^D = 10.08 \times 10^{-12}$
$g_{14} = 62 \times 10^{-8}$
$g_{25} = 170 \times 10^{-8}$

$g_{36} = 48 \times 10^{-8}$
$\epsilon_{22}{}^T = \dfrac{1}{\beta_{22}{}^T} = 10.0$
$\epsilon_{33}{}^T = \dfrac{1}{\beta_{33}{}^T} = 10.2$

For the mks system the elastic compliances are multiplied by 10, the piezoelectric constant g is multiplied by 3×10^5, and the dielectric constants are multiplied by 8.85×10^{-12}.

The only constant in the above equations that varies widely with temperature and field strength is the dielectric constant $\epsilon_{11}{}^T$ (the inverse of $\beta_{11}{}^T$). For low field strengths and frequencies above 1 kc, the dielectric constant as a function of temperature is shown by the dotted line of Fig. 11. It rises to very high values at the two Curie temperatures -18 and $+24$ deg cent. For high field strengths the dielectric constant, as shown by the solid line of Fig. 11, measured by the average slope of the hysteresis loop, becomes larger between the Curie points than it is at the Curie temperatures.

USEFUL CUTS IN ROCHELLE SALT. The cut most widely used is the X cut, which as shown by Fig. 9(*b*) is cut with its major face normal to the X axis. If a voltage is applied to this cut it shears so that the square changes into a rhombus. By cutting the crystal length 45° from the crystallographic Y and Z axis, a crystal is obtained which elongates along one direction and contracts along the width. This cut which is known as the 45° X cut is widely used in producing longitudinal vibrations. By combining two longitudinal crystals as shown by Fig. 9(*c*) a "bimorph" crystal is obtained which bends. This has a much lower frequency than a longitudinal crystal and is used in voice-frequency apparatus for picking up and reproducing sound. Figure 9(*d*) shows a combination of two X-cut crystals used to produce a twisting motion. The center faces of the two crystals form one set of electrodes and the two outside electrodes the other pair so that two opposing face shears are applied to the combination. This causes the whole crystal to twist and produces a torsional motion in the pair. Finally Fig. 9(*e*) shows two thin face shear X-cut crystals which, when they are clamped on three corners, produce a large motion at the fourth corner. All three of these bimorph type-crystals have been used in such devices as phonograph pickups, microphones, headphones, loudspeakers, surface-roughness analyzers, and light valves and have many other applications.

For a 45° X-cut crystal the equations applicable for the extension are

$$S_l = s_{22}'^D T_l + g_l \frac{D_x}{4\pi}$$
$$E_x = -g_l T_l + \beta_1{}^T D_x \tag{12}$$

where S_l is the strain along the length, T_l the stress applied along the length, g_l the effective piezoelectric constant for the 45° axis, and β^T the impermeability (inverse of the dielectric constant) which is measured when the crystal is free to move. In cgs units the above constants have the values

$$s_{22}'^D = 3.16 \times 10^{-12} \quad \text{cm}_2 \text{ per dyne}; \qquad g_l = 31 \times 10^{-8} = \frac{g_{14}}{2} \tag{13}$$

while the free dielectric constant, which is the inverse of β^T, has the value shown by Fig. 11 for low applied fields and for high fields (500 volts per centimeter). Equations (12) can be used to predict the action of the crystal under static conditions or at frequencies much lower than the resonant frequencies of the crystal. For example, if we wish to find the response of the crystal as a microphone, the second equation states that, for open-circuit conditions for which the charge on the surface, and hence the electrical displacement D_x, is zero, the potential generated for a given pressure (negative of the tension T_l) is

$$E_x = \frac{E}{l_t} = g_l T_l = 31 \times 10^{-8} \quad \text{(pressure in dynes per cm}^2\text{)} \tag{14}$$

Since the electrostatic unit of potential, the statvolt, is 300 volts, the volts generated per dyne per square centimeter pressure are

$$E_{\text{volts}} = 31 \times 10^{-8} \times 300 \times l_t \times p = 9.10 \times 10^{-5} \quad \text{volt per dyne per sq. cm. for a crystal 1 cm thick} \tag{15}$$

Since the voltage generated for a given pressure is directly proportional to the g_l constant, which is one-half the appropriate shear constant, eqs. (11) shows that a 45° Y-cut crystal, which will have a g_l piezoelectric constant equal to $\frac{1}{2} \times 170 \times 10^{-8} = 85 \times 10^{-8}$, will generate about 3 times the open-circuit voltage for the same pressure that a 45° X-cut crystal will. The 45° Y-cut has been used to some extent as a microphone and as a transducer in underwater sound equipment for transforming electrical into mechanical energy. When the crystal is used as a microphone working into a low impedance, the Y-cut will not deliver as much voltage as an X-cut crystal on account of the very low impedance (high capacity) of the X-cut crystal, but the voltage that it does deliver is not a function of the temperature as is the voltage of the 45° X-cut.

By eliminating D_x from eqs. (12), the strain S_l, which is the expansion per unit length, can be expressed in terms of the applied field as

$$S_l = \left[s_{22}{}^D + \frac{g_l{}^2 K_1{}^{LC}}{4\pi} \right] T_l + \frac{g_l \epsilon_1{}^T}{4\pi} E_x \tag{16}$$

In the absence of an external stress T_l the total free displacement of 1 volt applied is

$$d = S_l l = 1.03 \times 10^{-9} \frac{\epsilon_a^T}{4\pi} E \frac{l}{l_5} \tag{17}$$

This displacement as a function of the volts per inch applied is shown by Fig. 12 for several different temperatures. Outside the Curie region the displacement is much less since the dielectric constant ϵ_a^T is so much smaller, particularly for large fields.

When two crystals are glued together to form a bimorph unit it has been shown (see W. P. Mason, *Electromechanical Transducers and Wave Filters*, p. 214, Van Nostrand) that the displacement of the component longitudinal crystals is multiplied by the factor $3l/l_t$, where l is the length of the crystal and l_t the total thickness of the two elements. This is a method of enhancing the total displacement of the unit at the expense of a considerable lowering of the resonant frequency of the device. Since the dielectric constants of the two crystals glued together will be less than the free dielectric constant of Fig. 11 and will approach the dielectric constant of the clamped crystal shown by Fig. 13, the very

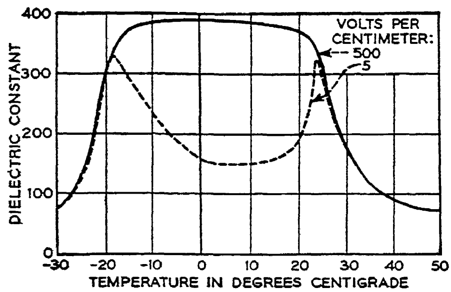


FIG. 12. Clamped Dielectric Constant of an X-cut Rochelle Salt Crystal as a Function of Temperature and Field Strength

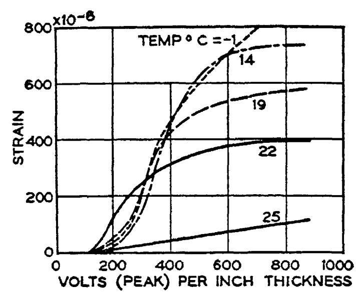


FIG. 13. Strain in Rochelle Salt as a Function of Temperature and Field Strength

large temperature and saturation effects noted for the free crystal will be considerably reduced for the bimorph type. However, the response may vary by a factor of 5 for a wide temperature range. A typical response in the ferroelectric range for a bender unit 1 1/2 in. long, 3/4 in. wide, and 0.040 in. thick is

33 volts	0.002 in.
77 volts	0.0045 in.
125 volts	0.006 in.
140 volts	0.0056 in.

The displacement for any other shape unit will vary in proportion to the factor $(l/l_t)^2$ and will be independent of the width.

When such units are used as voltage generators as in phonograph pickup devices, the mechanical impedance of the device is very considerably lowered over what would be obtained with a clamped longitudinal device. The response can be calculated by determining how much strain is generated by a given motion and calculating the voltage from eqs. (12). A typical unit 0.030 in. thick, 11/16 in. long, and 7/16 in. wide will give an output as high as 1 volt when played from a phonograph record. This response will be relatively independent of the temperature when the device is worked into the grid of a vacuum tube.

35. PROPERTIES OF AMMONIUM DIHYDROGEN PHOSPHATE (ADP)

Ammonium dihydrogen phosphate, which has been given the abbreviation ADP, is one member of four isomorphous salts whose dielectric properties were first investigated by Busch (Neue Seignette Elektrica, *Helv. Phys. Acta*, Vol. 11, No 3 [1938]). Two members of this group, namely, potassium dihydrogen phosphate and potassium dihydrogen arsenate, were found to have ferroelectric effects at 121 and 91 deg absolute temperature. Of these isomorphous crystals ADP was the crystal which had the largest piezoelectric coupling (about 30 per cent), and it was widely used during World War II as the transducing element for underwater sound projectors and hydrophones. It appears likely that, for devices that transform mechanical vibrations into electrical vibrations—phonograph pickups, microphones, etc.—ADP will give superior results to Rochelle salt and may eventually replace it for such applications. For devices that have to produce a large motion for a given voltage, however, Rochelle salt is still the only crystal that has a large enough d_{14} constant to be of interest.

ADP crystallizes in the tetragonal scalenohedral class with the habit shown by Fig. 14. The c or Z axis lies along the long direction of the crystal; this is an axis of fourfold alternating symmetry. The X and Y axes lie normal to the prism faces; they are axes of two-fold symmetry. Since the properties of crystals cut normal to these two surfaces are identical it is a matter of convention which is called X and which Y. The two diagonal axes, labeled P_1 and P_2, can be distinguished by piezoelectric tests, and P_1 has been taken as that axis along which a positive stress (tension) produces a positive charge at the positive (i.e., the upper) end of the Z axis. With the Z axis vertical and the P_1 axis toward the observer's right hand, the X axis has been taken as the axis that runs from front to back of the crystal and the Y axis the one that runs from left to right.

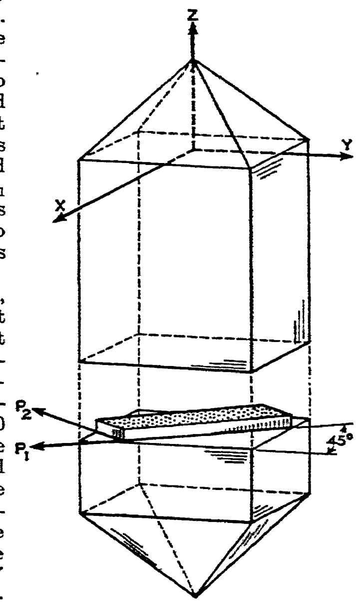


FIG. 14. 45° Z-cut ADP Crystal and Form of Natural Crystal

ADP, which has the chemical formula $NH_4H_2PO_4$, has no water of crystallization and hence will not dehydrate when the humidity becomes low. At about 93 per cent humidity, the crystal will deliquesce and will pick up water from the atmosphere. In practice it is necessary to keep the crystal in an atmosphere for which the humidity is 50 per cent or less since the water collected on the surface provides a leakage path across the crystal edges which becomes low enough to cause trouble for humidities above 50 per cent. Owing to the hydrogen bond system ADP also has a volume leakage which for a pure salt is shown by Fig. 15. For the Z-cut crystal having a dielectric constant of 15.7 this leakage will impair the response only for frequencies below 1 cycle per second. However, certain impurities introduced by the growing process can markedly decrease this resistivity, and for some applications it is necessary to specify a high resistivity. ADP can be taken up to 180 deg cent before it melts. However, ammonia is given off from the surface at temperatures above 100 deg cent, and since this impairs the adherence of the electrodes to the crystal surface it is desirable to keep the temperature of operation under 100 deg cent. The crystal, therefore, is useful under any likely ambient temperature conditions.

The piezoelectric equations for ADP take the form

$$\begin{aligned}
S_1 &= s_{11}T_1 + s_{12}T_2 + s_{13}T_3 & S_6 &= s_{66}{}^D T_6 + g_{36}\frac{D_z}{4\pi} \\
S_2 &= s_{12}T_1 + s_{11}T_2 + s_{13}T_3 & E_x &= \beta_1{}^T D_x - g_{14}T_4 \\
S_3 &= s_{13}T_1 + s_{13}T_2 + s_{33}T_3 & E_y &= \beta_1{}^T D_y - g_{14}T_5 \\
S_4 &= s_{44}{}^D T_4 + g_{14}\frac{D_x}{4\pi} & E_z &= \beta_3{}^T D_z - g_{36}T_6 \\
S_5 &= s_{44}{}^D T_5 + g_{14}\frac{D_y}{4\pi}
\end{aligned} \qquad (18)$$

where in cgs units the constants have the following values:

$$\begin{aligned}
s_{11} &= 1.74 \times 10^{-12} \text{ cm}^2 \text{ per dyne} & s_{44}{}^D &= 11.4 \times 10^{-12} \text{ cm}^2 \text{ per dyne} & \epsilon_1{}^T &= 59.0 \\
s_{12} &= 0.7 \times 10^{-12} & s_{66}{}^D &= 14.7 \times 10^{-12} & \epsilon_3{}^T &= 15.7 \\
s_{13} &= -1.1 \times 10^{-12} & g_{14} &= 1.06 \times 10^{-8} \\
s_{33} &= 4.35 \times 10^{-12} & g_{36} &= 118.5 \times 10^{-8}
\end{aligned} \qquad (19)$$

USEFUL CUTS FOR ADP CRYSTALS. Since the g_{36} constant is so much larger than the g_{14} constant in ADP it is obvious that most of the useful cuts will be those that are normal or nearly normal to the Z axis. A crystal cut normal to the Z axis will generate a face shear motion similar to that shown by Fig. 9(b) for Rochelle salt. Hence by cutting the length 45° from the X and Y crystallographic axes a longitudinal motion may be produced. The Z and the 45° Z cut are the principal ones used for ADP crystals. The Z cut has been used in generating face shearing modes and in the production of torsional crystals.

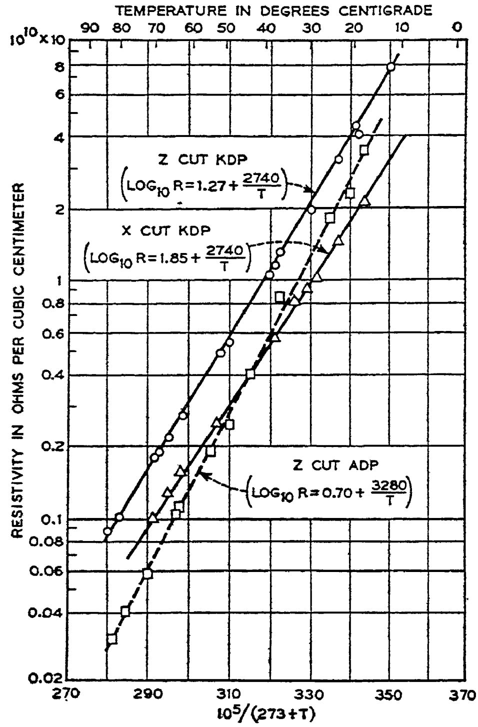


Fig. 15. Resistivity of ADP and KDP as a Function of Temperature

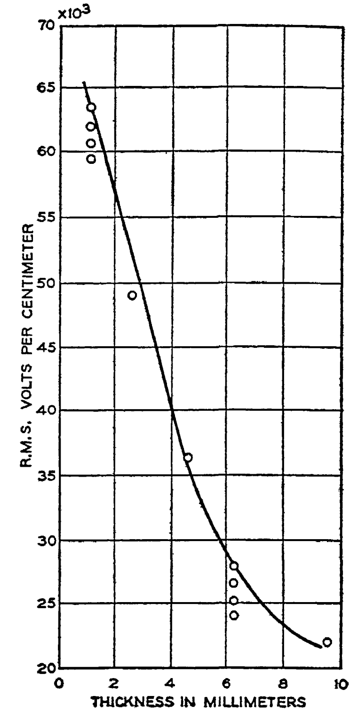


Fig. 16. Breakdown Voltage for an ADP Crystal as a Function of Thickness

The 45° Z-cut crystal has been used as the transducing element in underwater sound equipment and in microphones, in phonograph pickups, and in devices for transforming mechanical energy into electrical energy.

The equations of motion of a 45° Z cut take the form

$$S_l = s_{22}^{D\prime} T_l + g_l \frac{D_z}{4\pi} \qquad (20)$$

$$E_z = -g_l T_l + \beta_3^T D_z$$

where $g_l = g_{36}/2 = 59.2 \times 10^{-8}$; $s_{22}^D = 4.72 \times 10^{-12}$ cm² per dyne;

$$\epsilon_3^T = \frac{1}{\beta_3^T} = 15.7$$

When a crystal 1 cm thick is used as a voltage-generating device the number of volts generated on open circuit per dyne per square centimeter is

$$E_{\text{volts}} = 1.78 \times 10^{-4} \quad \text{volt} \qquad (21)$$

This is larger than for 45° X-cut Rochelle salt. On account of the lower dielectric constant this crystal has to be worked into a higher impedance than Rochelle salt to obtain the same output. Crystals of this sort are replacing Rochelle salt for applications such as microphones and phonograph pickups on account of their greater chemical stability and their ability to withstand wide temperature variations.

The electromechanical coupling factor of ADP is given by the formula

$$k = \frac{g_{36}}{2} \sqrt{\frac{\epsilon_3^T}{4\pi s_{22}'}} = 0.3 \qquad (22)$$

As can be seen from eq. (9) these crystals can convert electrical into mechanical energy, or vice versa, efficiently over a frequency range of

$$\frac{f_B}{f_A} = 1.36 \tag{23}$$

Considerable amounts of power can be transformed from electrical into high mechanical impedance systems. The crystal limitations are the breaking strain and the voltage gradient that the crystals will stand. Experiments with ADP crystals show that they will break if the strain exceeds from 4 to 10 × 10^{-4} cm per cm. The voltage gradient that they will stand before a voltage puncture occurs is a function of the thickness of the crystal. Figure 16 shows the voltage gradient that will produce a puncture on the average crystal. It can be shown that the particle velocity on the end of a quarter-wavelength or half-wavelength crystal is equal to

$$\dot{\xi} = vS_M \quad \text{or} \quad \dot{\xi} = 3.3 \times 10_4 S_M \tag{24}$$

where $\dot{\xi}$ is the particle velocity, v is the velocity of propagation, and S_M is the maximum strain that the crystal will suffer. This maximum strain occurs at the middle of a half-wavelength unit or at the glued joint of a quarter-wavelength unit. Since the crystal is stronger than most of the adhesives that can be used to attach it to high-mechanical-impedance solid materials, the half-wavelength crystal can be used to produce more power output than the quarter-wavelength unit. The two types of units are shown by Fig. 17.

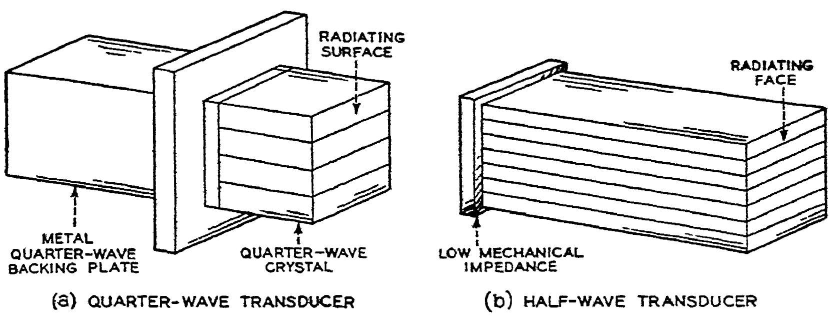


FIG. 17. Quarter- and Half-wave Transducers

The quarter-wavelength unit is glued to a heavy metal backing plate and radiates its energy from its free face. A half-wavelength unit, on the other hand, works into a low mechanical impedance on one end and radiates its energy from the other end. For a half-wavelength unit the maximum strain that the crystal can safely stand is 4 × 10^{-4}, so that the maximum particle velocity obtainable for an ADP crystal is about

$$\dot{\xi}_m = 132 \quad \text{cm per sec} \tag{25}$$

If the particle velocity is working into a high mechanical impedance such as the radiation impedance of water, which is R × 1.5 × 10^5 mechanical ohms per square centimeter, the energy radiated with this particle velocity is

$$E_{(\text{ergs/sec/sq/cm})} = \dot{\xi}_m{}^2 R = 2.6 \times 10^9 \quad \text{ergs per sec per sq cm} \tag{26}$$
$$= 260 \quad \text{watts per sq cm}$$

This power usually exceeds the power allowed by the voltage puncture limit (unless the crystals are made very thin) and this is the usual crystal limitation.

It can be shown that for a half-wavelength ADP crystal radiating into a mechanical load of R_M mechanical ohms per square centimeter (ratio of force in dynes per square centimeter to velocity in centimeters per second) the power radiated in watts per square centimeter is given in terms of the voltage gradient in volts per centimeter and the mechanical load R_M by the equation

$$\text{Power} = \frac{(E/l_t)^2}{24R_M} \quad \text{(half-wavelength radiator)} \tag{27}$$

For a quarter-wavelength unit, the power radiated is one-fourth of this for the same voltage gradient or

$$\text{Power} = \frac{(E/l_t)^2}{96R_M} \tag{28}$$

Hence if we know the limiting voltage gradients, the amount of power that the crystal will radiate before it punctures can be calculated. For example, if a half-wavelength radiator is working into the radiation impedance of water and is made up of crystals 0.25 cm thick, the crystals should withstand a voltage gradient of 50,000 rms volts per centimeter. From eq. (27) the power radiated should be 660 watts per square centimeter before dielectric breakdown occurs. Crystal fracture then becomes the limiting factor. In practice the limitation of power does not lie with the crystal but occurs in the medium if this is liquid or in the glued joint if a solid is used to transmit the power. A usual figure for continuous power is 5 watts per square centimeter. If the crystal is to radiate into a gas such as air, the limiting power is invariably determined by the limiting strain that the crystal can stand before it breaks. From eq. (28) the maximum amount of power that can be radiated into air by an ADP crystal is 0.075 watt per square centimeter since the radiation impedance of air is 43 ohms per square centimeter. A crystal is not an efficient means for exciting an air vibration.

Since the limiting particle velocity for an ADP crystal is about 130 cm per sec, a crystal cannot be used directly to produce very high strains in metals or terminal velocities approaching the speed of sound. If, however, a crystal mosaic is glued to a metal rod tapered exponentially like a horn, a very high strain and a very high terminal velocity can be produced at the small end. Figure 18 shows a construction proposed by the writer for

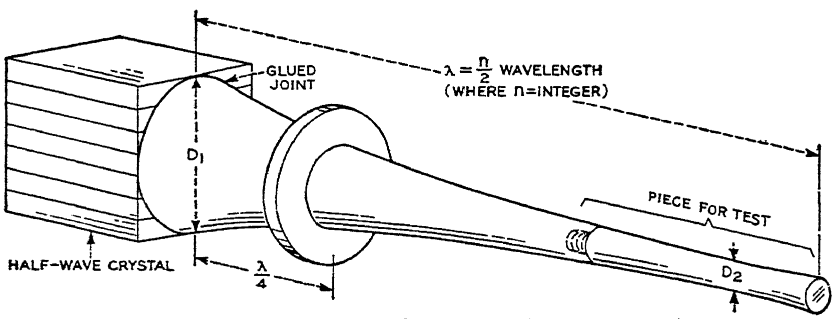


FIG. 18. Mechanical Horn for Producing a Large Strain in a Metal Sample

testing fatigue in metals. A crystal mosaic several inches in cross-section is glued to a steel rod which tapers from the crystal area down to a thickness of 0.05 in., after which it increases in diameter. The taper is an exponential function of the length and must satisfy the relation

$$T \leqq \frac{2\pi f}{v_s} \tag{29}$$

where the taper T is determined by the equation for the area

$$S = S_0 e^{-2Tl} \tag{30}$$

where f is the resonant frequency of the crystal, and v_s the velocity of sound in the steel. If the total length of the steel piece is made an integral number of half wavelengths of the frequency, the glued joint will come at a loop of the motion and will not be appreciably strained. The whole system will act as a resonant system and will produce a considerable motion for small applied voltages. The steel section adjacent to the crystal will have the same particle velocity as the crystal surface. Now it can be shown that the strain in the bar of uniform section at the nodal point (point of maximum strain) is equal to

$$S = \frac{\dot{\xi}}{v_s} \tag{31}$$

where v_s is the velocity of propagation of the wave in the steel and $\dot{\xi}$ the particle velocity at the surface. The effect of the tapered section is to increase the particle velocity in inverse proportion to the diameter. Hence if the diameter decreases from 2 in. to 0.05 in. the velocity is multiplied by a factor of 40 and the strain at the nodal point is equal to

$$S_s = 40 \times \frac{v_c}{v_s} S \tag{32}$$

where S_s is the strain in the steel, S_c the strain in the crystal, v_c the velocity of propagation in the crystal (3.3×10^5 cm per sec), and v_s the velocity of propagation in the steel (about

5.1×10^5 cm per sec). Hence, for a strain of 4×10^{-4} in the crystal, a strain of 0.01 can be generated in the steel. This is sufficient to cause plastic deformation in the steel, and by gradually increasing the drive on the crystal the fatigue properties of the steel can be investigated at a high rate of strain and of velocity.

This same system can be used to produce a high particle velocity on the small end of the steel bar. The only limitation is the strain that the metal will stand. Other uses appear to be delivering a large amount of power for a small area. By mounting a torsional crystal on the large end of the bar a torsional vibration can be given to the bar and the properties of the material under shearing strain can be tested. An ADP crystal can be made to vibrate in torsion by using the electrode system shown by Fig. 19. The inside

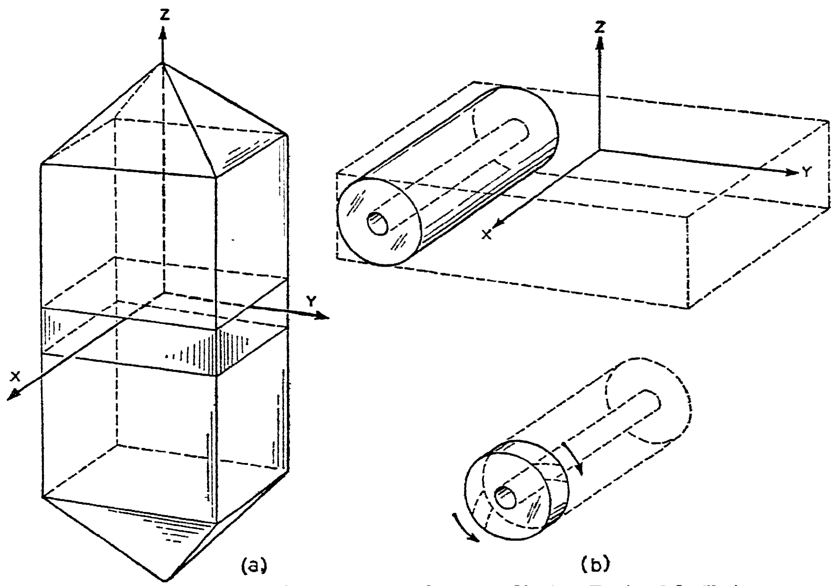


FIG. 19. Method for Cutting an ADP Crystal to Obtain a Torsional Oscillation

surface is covered by one electrode while the two outside electrodes, each of which covers a 90° segment, are connected together and form the other electrode. The centers of the two outside electrodes are normal to the Z axis, and the field is directed out from the center for both electrodes as shown by Fig. 19(b), thus producing a shearing motion for one segment and the opposite shear on the other segment so that the whole crystal is given a torsional motion.

SECTION 14

OPTICS

GEOMETRICAL OPTICS

By D. W. Epstein

VISION

By Kenneth N. Ogle

ELECTRON OPTICS

By D. W. Epstein

OPTICS

GEOMETRICAL OPTICS

By D. W. Epstein

Geometrical optics is mainly concerned with the geometrical relations of the propagation of radiant energy.

Neglecting quantum effects, the propagation of radiant energy is governed by Maxwell's equations. Fermat's principle of least time, which is the fundamental law governing the propagation of "rays" in geometrical optics, follows from Maxwell's equations if the wavelength of the radiation is allowed to approach zero. Although geometrical optics applies strictly only to the propagation of radiation of zero wavelength, it provides a very good and extremely useful approximation to any case where the wavelength is negligibly small in comparison with the smallest linear dimension of the apparatus.

Fermat's principle of least time states that the path traversed by light in passing between two points is that which will take the least time. The general law expressed by Fermat's principle is also known as the law of extreme path. It is stated mathematically as

$$\delta \int_A^B N\,dS = 0, \tag{1}$$

$N = \dfrac{C}{V}$ = index of refraction.

C = velocity of radiation in vacuum.

V = velocity of radiation in medium.

dS = element of path length.

The product of index of refraction and path length, $N\,dS$, is known as the optical length of a "ray of light" or optical distance. Equation (1) states that a light ray going from point A to point B will always choose that path which will make the optical distance an extremum (generally a minimum but sometimes a maximum) with respect to all neighboring paths for rays of the same frequency. The laws of linear propagation, reflection, and refraction may be deduced from Fermat's principle.

Although the frequency is a constant for a given radiation, and its wavelength varies with the medium traversed, it has become customary to specify radiation, especially visible radiation, by its wavelength in vacuum. This is due to the fact that fundamental measurements yield wavelength rather than frequency. The units of wavelength commonly used in optics are:

$$\begin{aligned} \text{micron } (\mu) &= 10^{-6} \text{ meter} \\ \text{millimicron } (\text{m}\mu) &= 10^{-9} \text{ meter} \\ \text{angstrom unit (A)} &= 10^{-10} \text{ meter} \end{aligned}$$

The visible spectrum extends from about 0.39 to 0.75 μ; or 390 to 750 mμ; or 3900 to 7500 A; or from about 4.0×10^{14} to 7.7×10^{14} cycles per second.

1. REFLECTION AND REFRACTION

When a beam of radiant flux or luminous flux strikes a boundary separating two homogeneous isotropic media, it is in general partly reflected and partly refracted. If the boundary is smooth (relative to the wavelength of radiation), the following simple laws of refraction and reflection apply (see Fig. 1).

LAW OF REFRACTION OR SNELL'S LAW. The ratio of the sine of the angle of incidence to the sine of the angle of refraction is constant depending only on the indices of refraction of the two media; i.e.,

$$N_1 \sin I_1 = N_2 \sin I_2 \tag{2}$$

The incident ray, the refracted ray, and the normal to the surface at the point of incidence all lie in the same plane.

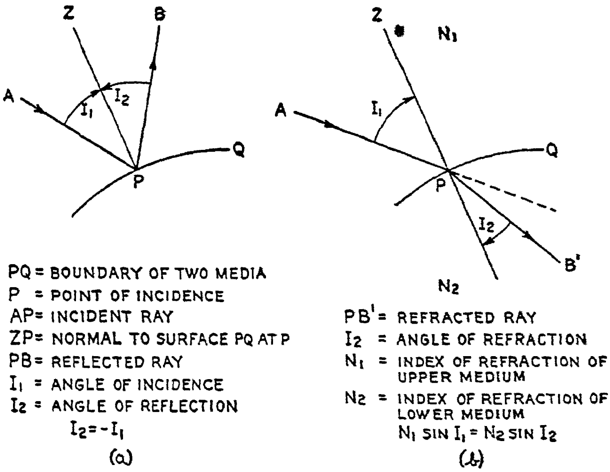


FIG. 1. Reflection and Refraction of Light Ray

LAW OF REFLECTION. The angle of reflection is equal to the angle of incidence; i.e.,

$$I_2 = -I_1 \tag{3}$$

The incident ray, the reflected ray, and the normal to the surface at the point of incidence all lie in the same plane.

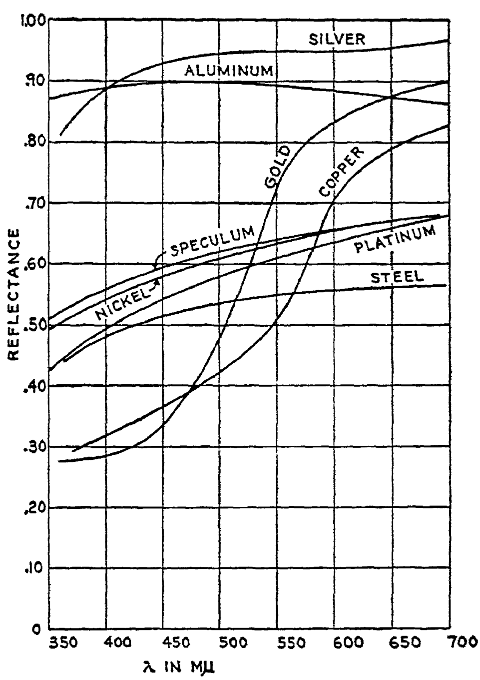


FIG. 2. Reflectance of Some Metals as a Function of Wavelength

Reflection from smooth surfaces is called *specular* or regular reflection. If the boundary is between a transparent dielectric (air, glass) and a metal, then, in general, most of the incident flux is reflected. Figure 2 gives the specular reflectance (ratio of reflected to incident flux) of various metals as a function of wavelength.

The specular reflectance r_f at the boundary of two transparent dielectrics is specified for normal incidence by Fresnel's equation:

$$r_f = \frac{(N_2 - N_1)^2}{(N_2 + N_1)^2} \tag{4}$$

where N_1 and N_2 are the indices of refraction of the two media.

Figure 3 shows the reflectance for unpolarized light as a function of angle of incidence for light traveling from a lower-index medium into a higher-index medium ($N_2 > N_1$). It is seen that, as the angle of incidence is increased, the reflectance rises gradually at first, and then rapidly, until it becomes unity at the angle of incidence of 90°. The remainder of the incident flux enters the medium of index N_2 at the angle of refraction I_2 given by Snell's law, eq. (2).

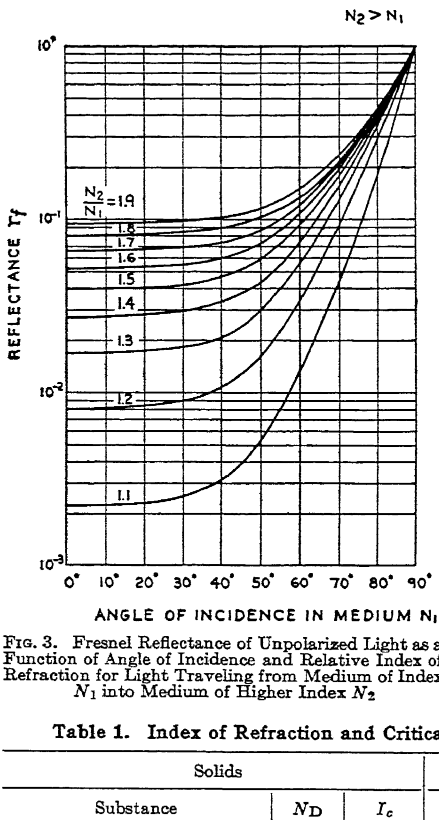


FIG. 3. Fresnel Reflectance of Unpolarized Light as a Function of Angle of Incidence and Relative Index of Refraction for Light Traveling from Medium of Index N_1 into Medium of Higher Index N_2

The Fresnel reflection can be greatly reduced by means of a transition layer between the two transparent dielectrics. For normal incidence and for a particular wavelength, the Fresnel reflection can be reduced to zero if, as in transmission lines, the index of refraction of the transition layer is $N = \sqrt{N_1 N_2}$ and the thickness of the layer is $\lambda/4$.

For light traveling from a higher-index medium into a lower-index medium ($N_2 < N_1$), a similar behavior is observed up to a certain critical angle of incidence, the critical angle being determined by the condition that

$$\sin I_1 = \sin I_C = \frac{N_2}{N_1} \tag{5}$$

since at this angle of incidence the angle of refraction is 90° ($\sin I_2 = 1$). For angles of incidence greater than the critical angle, the beam is totally reflected into the initial medium. Table 1 gives the index of refraction and critical angle of incidence relative to air for some solids and liquids.

Table 1. Index of Refraction and Critical Angle of Substances Relative to Air

Solids			Liquids, 15 deg cent		
Substance	N_D	I_c	Substance	N_D	I_c
Magnesium fluoride	1.38	46° 26′	Water	1.334	48° 33′
Quartz (fused)	1.458	43° 18′	Ether	1.357	47° 28′
Pyrex (glass)	1.474	42° 43′	Alcohol (ethyl)	1.364	47° 9′
Methyl methacrylate	1.49	42° 9′	Glycerine	1.471	42° 50′
Potassium chloride	1.49	42° 9′	Carbon bisulfide	1.630	37° 51′
Canada balsam	1.526	40° 57′	Methylene iodide	1.732	35° 16′
Sodium chloride	1.544	40° 22′	50% Methylene iodide & 50% phosphorus	1.929	31° 14′
Polystyrene	1.59	38° 58′			
Willemite	1.71	35° 47′			
Calcium tungstate	1.92	31° 23′			
Zinc oxide	2.05	29° 12′			
Zinc sulfide	2.37	24° 57′			
Diamond	2.417	24° 26′			
Cadmium sulfide	2.52	23° 23′			

DIFFUSE REFLECTION. If the boundary between two media is rough, specular reflection can be considered to exist at a great number of small smooth areas oriented in various directions, and the reflected energy is distributed over a wide range of angles. In general, practical surfaces (boundaries) reflect partly specularly and partly diffusely.

Figure 4 illustrates this. Pure specular reflection is shown in (*a*); part specular and part diffuse, such as would occur at most matte surfaces, is shown in (*b*); pure or ideal diffuse reflection is shown in (*c*). In perfectly diffuse reflection, the flux reflected per unit solid angle is proportional to the cosine of the angle measured from the normal to the surface. This statement is known as *Lambert's law*.

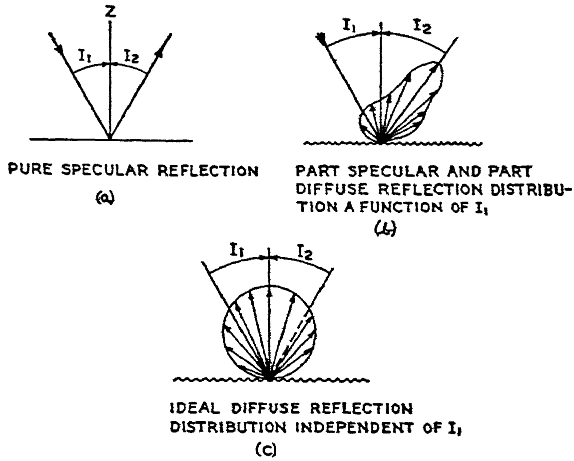


FIG. 4. Specular and Diffuse Reflection

DISPERSION. The variation of the refractive index of a substance with the wavelength (color) of the transmitted light is termed dispersion. The index of refraction of most glasses varies with wavelength in a manner which may be approximated by the dispersion formula of Cauchy:

$$N = a + \frac{b}{\lambda^2}$$

The refractive index of optical glass is generally measured with certain definite wavelengths or lines of the spectrum. It has become customary to use the spectral lines A′ (0.7665 μ); C (0.6563 μ); D (0.5893 μ); F (0.4861 μ); G′ (0.4341 μ). The differences in refractive index $N_D - N_C$, $N_F - N_D$, $N_{G'} - N_F$, $N_D - N_{A'}$, and $N_F - N_C$ are taken as a measure of the dispersion of the glass in the different parts of the spectrum. If a glass is specified by only one index of refraction N_D is generally meant. A quantity related to dispersion which is in very general use utilizes the F (blue), D (yellow), and C (red) portions of the spectrum and is designated by:

$$V = \frac{N_D - 1}{N_F - N_C} \tag{6}$$

The V values, $N_{A'}$, N_C, N_D, N_F, and $N_{G'}$, for some Bausch & Lomb Optical Company glasses are given in Table 2.

Table 2. Indices of Refraction and V-number of Some Optical Glasses Made by Bausch and Lomb Optical Co.

Type of Glass	$N_{A'}$ 766.5 mμ	N_C 656.3 mμ	N_D 589.3 mμ	N_F 486.1 mμ	$N_{G'}$ 434.1 mμ	$V = \frac{N_D - 1}{N_F - N_C}$
Borosilicate Crown BSC-1	1.50578	1.50860	*1.51100*	1.51665	1.52114	63.5
Crown C-1	1.51729	1.52036	*1.52300*	1.52929	1.53435	58.6
Light Barium Crown LBC-1	1.53529	1.53842	*1.54110*	1.54746	1.55257	59.9
Dense Barium Crown DBC-1	1.60439	1.60793	*1.61100*	1.61832	1.62421	58.8
Crown Flint CF-1	1.52217	1.52560	*1.52860*	1.53584	1.54178	51.6
Light Barium Flint LBF-1	1.58110	1.58479	*1.58800*	1.59580	1.60212	53.4
Extra Light Flint ELF-1	1.55086	1.55495	*1.55850*	1.56722	1.57447	45.5
Barium Flint BF-2	1.59682	1.60130	*1.60530*	1.61518	1.62345	43.6
Light Flint LF-1	1.56425	1.56861	*1.57250*	1.58208	1.59011	42.5
Dense Flint DF-2	1.60684	1.61218	*1.61700*	1.62904	1.63929	36.6
Dense Barium Flint DBF-1	1.60731	1.61242	*1.61700*	1.62843	1.63811	38.5
Extra Dense Flint EDF-3	1.70555	1.71309	*1.72000*	1.73766	1.75304	29.3

REFRACTION AND REFLECTION AT SPHERICAL SURFACES. Because of the relative ease of manufacture, most optical systems consist of a series of spherical surfaces, a plane being the limiting case of a spherical surface of infinite radius of curvature. By determining the plane containing the ray to be traced and the center of curvature, any problem of refraction or reflection at a spherical surface may be reduced to a problem in plane trigonometry.

SIGN CONVENTION. Referring to Fig. 5, let the paper represent the plane of incidence containing A_1Q, the ray to be traced through the refracting surface, and C the center of curvature of the spherical surface. The signs of quantities in optical calculations have not been standardized, but the sign convention indicated in Fig. 5 is very widely used.

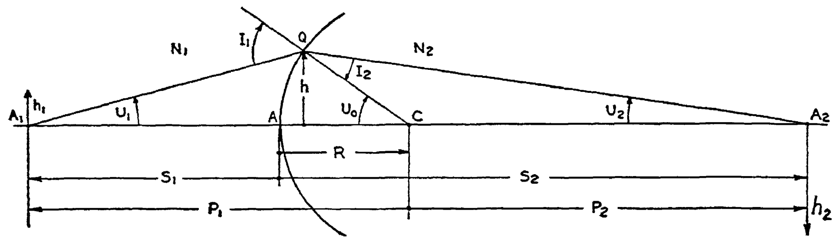


Fig. 5. Refraction at a Spherical Surface

Distances measured to the right of the pole A (or center of curvature C) are positive; those to the left of the pole A are negative. Distances above the axis are positive, and those below the axis are negative. Angles shown clockwise are positive; those shown counterclockwise are negative.

EXACT RAY TRACING EQUATIONS. Using the above sign convention, the standard ray tracing equations given below follow from the law of refraction and plane trigonometry. It is assumed that the quantities N_1, N_2, R, and S_1 (or P_1) are given and that U_1 is assigned an arbitrary value for each ray. The problem is to find I_1, I_2, U_2, and S_2 (or P_2).

$$\sin I_1 = \frac{S_1 - R}{R} \sin U_1 = \frac{P_1}{R} \sin U_1 \tag{7}$$

$$\sin I_2 = \frac{N_1}{N_2} \sin I_1 \tag{8}$$

$$U_2 = U_1 + I_1 - I_2 \tag{9}$$

$$S_2 - R = P_2 = R\,\frac{\sin I_2}{\sin U_2} \tag{10}$$

In dealing with a complete optical system with many surfaces the results of one surface are taken as the initial data for the next surface. The following equations which may also be derived with the aid of Fig. 5 have been found very useful in optical calculations:

$$P_2 N_2 \sin U_2 = P_1 N_1 \sin U_1 \tag{11}$$

$$U_0 = U_1 + I_1 = U_2 + I_2 \tag{12}$$

$$h = R \sin U_0 \tag{13}$$

$$\frac{N_1 \cos I_1}{P_2} - \frac{N_2 \cos I_2}{P_1} = \frac{N_2 \cos I_2 - N_1 \cos I_1}{R} \cos U_0 \tag{14}$$

Since at reflection $I_2 = -I_1$, $\sin I_2 = -\sin I_1$, the law of reflection may be treated mathematically as a particular case of the law of refraction, i.e., where $N_2 = -N_1$. Hence the refraction formulas given above apply to the case of reflection at a spherical surface by simply letting $N_2 = -N_1$. Thus for a spherical mirror, the above equations become (Fig. 6):

$$\sin I_1 = \frac{P_1}{R} \sin U_1 \tag{15}$$

$$\sin I_2 = -\sin I_1 \tag{16}$$

$$U_2 = U_1 - 2I_2 \tag{17}$$

$$P_2 = R\,\frac{\sin I_2}{\sin U_2} \tag{18}$$

$$P_2 \sin U_2 = -P_1 \sin U_1 \tag{19}$$

$$\frac{1}{P_2} + \frac{1}{P_1} = -\frac{2 \cos U_0}{R} \tag{20}$$

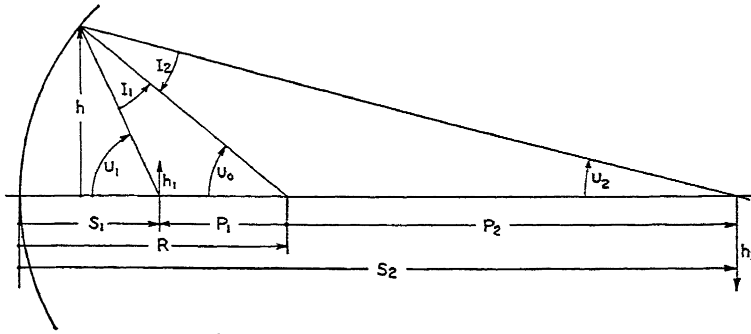


FIG. 6. Reflection at a Spherical Surface

PARAXIAL FORMULAS. The exact trigonometrical formulas just given are transcendental and are therefore extremely difficult to manipulate. Great simplification is obtained if the equations are restricted to paraxial rays, i.e., rays that make small angles with the optical axis and with the normals to the refracting and reflecting surfaces. The paraxial formulas given below are deduced from the exact formulas by replacing the sines of the angles by the angles and the cosines by unity. Paraxial quantities will be indicated by the corresponding lower-case letter:

$$i_1 = \frac{s_1 - R}{R} u_1 = \frac{p_1}{R} u_1 \tag{21}$$

$$i_2 = \frac{N_1}{N_2} i_1 \tag{22}$$

$$u_2 = u_1 + i_1 - i_2 \tag{23}$$

$$s_2 - R = p_2 = R \frac{i_2}{u_2} \tag{24}$$

$$p_2 N_2 u_2 = p_1 N_1 u_1 \tag{25}$$

$$\frac{N_1}{p_2} - \frac{N_2}{p_1} = \frac{N_2 - N_1}{R} \tag{26}$$

$$\frac{N_2}{s_2} - \frac{N_1}{s_1} = \frac{N_2 - N_1}{R} \tag{27}$$

The corresponding equations for a spherical mirror are:

$$i_1 = \frac{p_1}{R} u_1 \tag{28}$$

$$i_2 = -i_1 \tag{29}$$

$$u_2 = u_1 - 2i_2 \tag{30}$$

$$p_2 = R \frac{i_2}{u_2} \tag{31}$$

$$p_2 u_2 = -p_1 u_1 \tag{32}$$

$$\frac{1}{p_2} + \frac{1}{p_1} = -\frac{2}{R} \tag{33}$$

$$\frac{1}{s_2} + \frac{1}{s_1} = \frac{2}{R} \tag{34}$$

Paraxial relations are linear and therefore are easily manipulated algebraically. They are very useful for determining the approximate focusing action of an optical system.

If the object is at a very great distance, i.e., $1/s_1 = 0$, the image is located at the second focal point at the distance

$$f_2 = \frac{N_2}{N_2 - 1} R \tag{35}$$

to the right of the apex A. If the image is located at a very great distance, i.e., $1/s_2 = 0$, then the object is located at the first focal point at the distance

$$f_1 = -\frac{N_1}{N_2 - N_1} R \tag{36}$$

to the left of the apex A. For a spherical mirror

$$f_2 = f_1 = \frac{R}{2} \tag{37}$$

and both focal points coincide. It should be noted that in general

$$\frac{f_2}{f_1} = -\frac{N_2}{N_1} \tag{38}$$

This relation applies to any optical system where N_2 is the index of refraction of the image space and N_1 that of the object space.

MAGNIFICATION. The above relations suffice to locate an image of an object, but they do not explicitly give any information about the sizes of object and image. As is

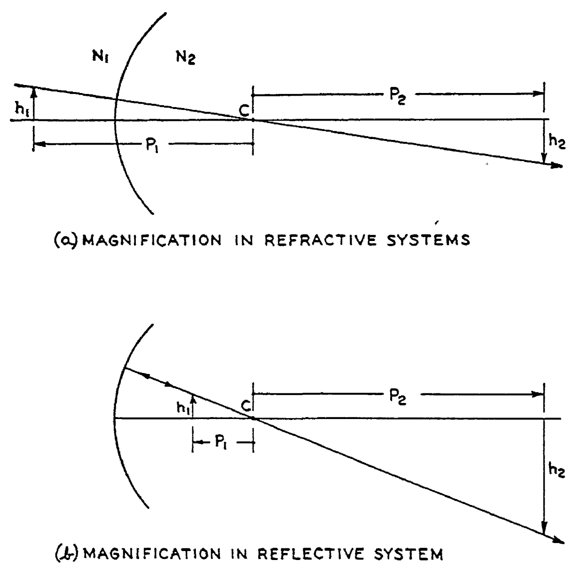


FIG. 7. Magnification in Optical Systems

seen from the geometry of Figs. 7(*a*) and 7(*b*), the *lateral magnification* which is defined as the ratio of image height h_2 to object height h_1 is

$$m = \frac{h_2}{h_1} = \frac{P_2}{P_1} \tag{39}$$

It is to be noted that m is negative for an inverted image (as in Fig. 7) and positive for an erect image. Inserting (39) into (11) there result the very important relations

$$h_2 N_2 \sin U_2 = h_1 N_1 \sin U_1 \tag{40}$$

$$m = \frac{N_1 \sin U_1}{N_2 \sin U_2} \tag{41}$$

Equation (40) or (41) is known as *Abbe's sine condition* and applies to any number of refracting and reflecting coaxial surfaces (N_1 index of object space, N_2 index of image space). The corresponding paraxial equations

$$h_2 N_2 u_2 = h_1 N_1 u_1 \tag{42}$$

$$m = \frac{N_1 u_1}{N_2 u_2} \tag{43}$$

are known as *Lagrange's theorem*.

The longitudinal magnification along the axis is

$$m_s = \frac{\Delta s_2}{\Delta s_1} = m^2 \tag{44}$$

The angular magnification is

$$m_u = \frac{u_2}{u_1} = \frac{1}{m} \tag{45}$$

The relation between these magnifications is

$$m = m_s m_u \tag{46}$$

2. LENSES

By applying the above paraxial relations to a lens in air, i.e., two refracting surfaces, as shown in Fig. 8, it may be shown that the two focal points (F_1, F_2) and the two principal points (H_1, H_2) completely determine the paraxial focusing characteristics of the lens.

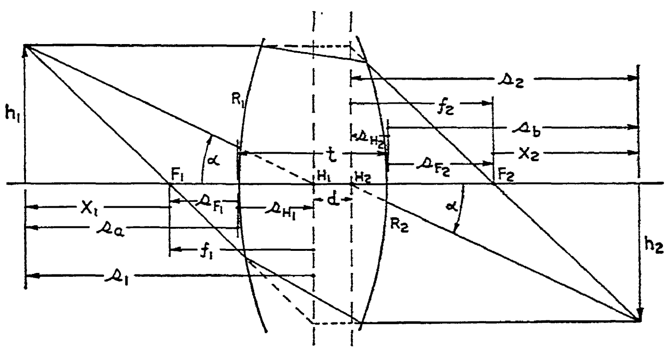


Fig. 8. Focusing Characteristics of Thick Lens in Air

The first focal point, F_1, may be considered as that object point which is focused at infinity in the image space so that all paraxial rays passing through the point F_1 are parallel to the axis in the image space. The second focal point, F_2, may be considered the image of an object at an infinite distance from the lens in the object space so that all paraxial rays which are parallel to the axis in the object space will pass through F_2 in the image space.

The distances indicated by f_1 and f_2 in Fig. 8 are known as the *principal* or *equivalent* focal lengths and are given by

$$f_2 = \frac{NR_1R_2}{(N-1)[N(R_2 - R_1) + (N-1)t]} = -f_1 \tag{47}$$

The positions of the focal points measured from the refracting surfaces are given by

$$s_{F_1} = -f_2\left(1 + \frac{N-1}{N}\frac{t}{R_2}\right) \tag{48}$$

$$s_{F_2} = f_2\left(1 - \frac{N-1}{N}\frac{t}{R_1}\right) \tag{49}$$

The focal distances s_{F_1} and s_{F_2} are known as the front focal length (F.F.L.) and back focal length (B.F.L.) respectively. These focal lengths are easily measured by imaging an object at a great distance from the lens and noting the distance between image and nearest surface of the lens and then turning the lens around and repeating the operation. The positions of the principal points H_1 and H_2 measured from the refracting surfaces are given by

$$s_{H_1} = -f_2\frac{N-1}{N}\frac{t}{R_2} \tag{50}$$

$$s_{H_2} = -f_2\frac{N-1}{N}\frac{t}{R_1} \tag{51}$$

The distance between the principal points is

$$d = \frac{N-1}{N} t \left[1 + \frac{N-1}{N} \frac{tf_2}{R_1R_2} \right] \tag{52}$$

From (50) and (51) it is seen that

$$\frac{s_{H_1}}{s_{H_2}} = \frac{R_1}{R_2} \tag{53}$$

In many practical cases the lens thickness, t, is numerically small in comparison with either R_1, R_2, or $(R_2 - R_1)$, and then

$$s_{H_1} = \frac{R_1 t}{N(R_2 - R_1)} \qquad s_{H_2} = \frac{R_2 t}{N(R_2 - R_1)} \tag{54}$$

$$d = \frac{N-1}{N} t \tag{55}$$

For a biconvex lens with $R_2 = -R_1(R_1$ positive)

$$f_2 = \frac{R_1}{2(N-1)} \qquad s_{H_1} = \frac{t}{2N} \qquad s_{H_2} = -\frac{t}{2N}$$

If $R_2 \neq R_1$, both principal planes move toward the surface with the higher curvature, and when $R_2 = \pm\infty$, i.e., for a plano-convex lens,

$$s_{H_1} = 0 \qquad s_{H_2} = -\frac{t}{N}$$

The positions of the focal and principal points of an optical system having been determined, its focusing action may be calculated with the aid of the following formulas (see Fig. 8).

$$X_1X_2 = f_1f_2 = -f_2^2 \tag{56}$$

$$\frac{1}{s_2} - \frac{1}{s_1} = \frac{1}{f_2} \tag{57}$$

$$m = \frac{h_2}{h_1} = -\frac{f_1}{X_1} = -\frac{X_2}{f_2} = \frac{s_2}{s_1} \tag{58}$$

$$s_2 = -f_2(m-1) \tag{59}$$

$$s_1 = -f_2 \frac{(m-1)}{m} \tag{60}$$

THIN LENSES. A lens of negligible thickness relative to its focal length is known as a thin lens. For a thin lens, $s_{H_1} = s_{H_2} = d = t = 0$, and great simplification in computation results. For a thin lens the distances s_1, s_2, f_1, and f_2 in eqs. (56) to (60) are measured from the (center of) lens, and the distance between object and image is

$$s_2 - s_1 = -f_2 \frac{(m-1)^2}{m} \tag{61}$$

COMPOUND LENSES. Because of the necessity for correcting aberrations most lenses are compound; i.e., they consist of several lenses having a common axis. Some of the surfaces of adjacent lenses may be in contact and others are separated by air spaces. The focal length of two thin lenses in contact of focal lengths f_{2a} and f_{2b} is

$$f_2 = \frac{f_{2a}f_{2b}}{f_{2a} + f_{2b}} \tag{62}$$

The power of a lens measured in diopters is defined as $P = 1/f$, where f is measured in meters. Thus the power of two thin lenses in contact is

$$P = \frac{1}{f_{2a}} + \frac{1}{f_{2b}} = P_a + P_b \tag{63}$$

If the two thin lenses are separated by an air space of thickness t, the resulting lens is effectively a thick lens and the focal length of the combination is

$$f_2 = \frac{f_{2a}f_{2b}}{f_{2a} + f_{2b} - t} = -f_1 \tag{64}$$

and the power is

$$P = P_a + P_b - tP_aP_b \tag{65}$$

The principal points and focal points are located at (see Fig. 9)

$$s_{H_1} = \frac{f_{2a}t}{f_{2a} + f_{2b} - t} \qquad s_{H_2} = -\frac{f_{2b}t}{f_{2a} + f_{2b} - t} \tag{66}$$

$$s_{F_1} = \frac{f_{2a}(f_{2b} - t)}{f_{2a} + f_{2b} - t} \qquad s_{F_2} = \frac{f_{2b}(f_{2a} - t)}{f_{2a} + f_{2b} - t} \tag{67}$$

The distance between the principal points is

$$d = -\frac{t^2}{f_{2a} + f_{2b} - t} \tag{68}$$

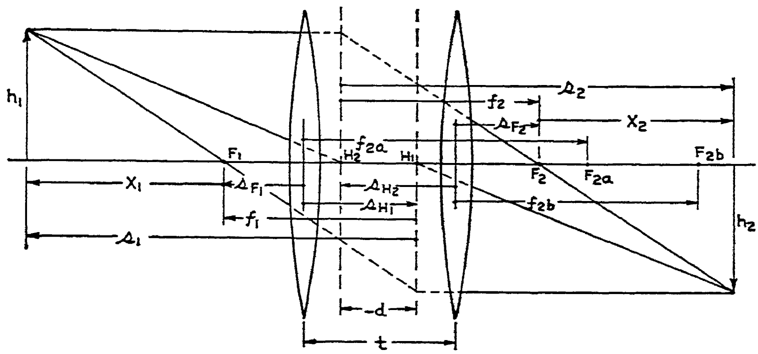


FIG. 9. Focusing Characteristics of Two Thin Lenses Separated by an Air Space

There are three ways of compounding two lenses.

1. Both lenses are converging; i.e., $f_{2a} > 0$ and $f_{2b} > 0$. In this case the focal length of the combination is a minimum and the power is a maximum when $t = 0$. The focal length increases as t is increased until when $t = f_{2a} + f_{2b}$ the focal length is infinite and the system is afocal or telescopic. Objectives, magnifiers, and eyepieces are generally compounded of two such lenses with $t < f_{2a} + f_{2b}$. The focal length of the combination is negative for $t > f_{2a} + f_{2b}$, and such a combination forms the compound microscope.

2. Both lenses are diverging; i.e., $f_{2a} < 0$ and $f_{2b} < 0$. The focal length of the combination is negative and decreases in absolute values as t is increased.

3. *One lens is converging* $f_{2a} > 0$ *and one diverging* $f_{2b} < 0$. If $|f_{2b}| < |f_{2a}|$ the focal length of the combination is negative for values of $t < |f_{2a} + f_{2b}|$, it is infinite for $t = |f_{2a} + f_{2b}|$; this is the optical system of the Galilean telescope which produces an erect image. The focal length is positive for $t > |f_{2a} + f_{2b}|$ and corresponds to the optical system of a telephoto lens. If $|f_{2b}| > |f_{2a}|$ the focal length is always positive.

ABERRATIONS. Actual imagery with practical lenses departs from the paraxial or first-order imagery discussed above. These departures are known as aberrations. There are seven independent third-order aberrations.

Spherical aberration or aperture defect is illustrated in Fig. 10. In the presence of spherical aberration the rays from a point object on the axis do not recombine to form a

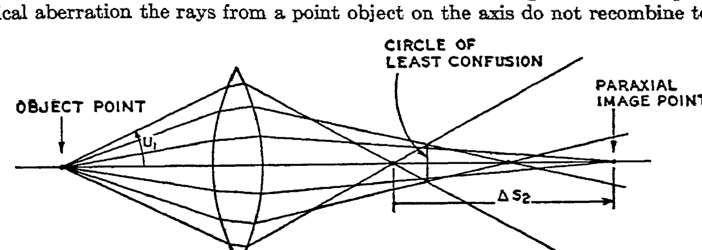


FIG. 10. Positive or Undercorrected Spherical Aberration

point image as required by paraxial theory. Spherical aberration is generally measured by the axial intersection distance, ΔS_2, which varies approximately as the cube of the convergence angle U_1. Any simple convex lens has positive spherical aberration and is corrected by compounding it with a concave lens which has negative spherical aberration.

In a single lens the spherical aberration is minimized if the deviation is equally divided between the two surfaces of the lens.

Coma is an extra-axial aberration; i.e., it affects image points not on the axis of the lens. It may be considered lateral spherical aberration for an extra-axial point. If coma is present a point object is imaged into a comet-shaped figure. It increases directly with the distance of the object point from the axis and as the square of the aperture of the lens. Coma, unlike spherical aberration, may be eliminated from a single thin lens for one object distance; however, a lens for zero coma is not the same as one for minimum spherical aberration. Coma may be eliminated by a stop of the proper aperture located at the proper distance from the lens. A lens corrected for spherical aberration and coma for a given object distance is called an *aplanatic lens*. If spherical aberration is absent the condition for freedom of coma is given by Abbe's sine condition, eq. (40) or (41).

Astigmatism, like coma, affects only extra-axial points. When astigmatism is present each object point has two images, one behind the other. These images are short lines at right angles to each other. One is called the sagittal or radial astigmatic line since it is in a plane containing the axis of the system. The other is called the tangential or transverse astigmatic line since it is at right angles to a line drawn from the lens axis. Midway the two line images the cross-section of the beam is a circle of least confusion. The astigmatism is positive if the sagittal focus is farther from the lens than the tangential focus. Otherwise it is negative. Astigmatism increases as the square of the distance of the object point from the axis and directly as the aperture.

Astigmatism and *curvature of the field* are generally present together, and the sagittal and tangential foci lie on curved surfaces as illustrated in Fig. 11. If only curvature of

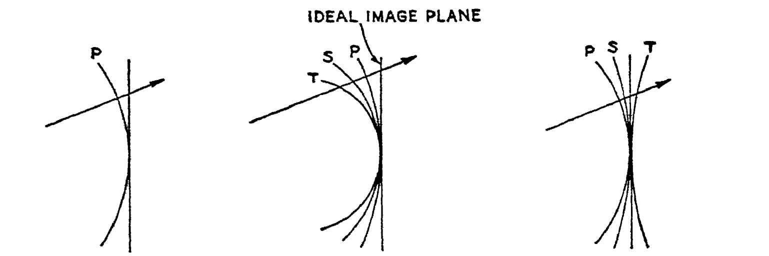


FIG. 11. Astigmatism and Curvature of Field

the field is present the image points lie on a spherical surface known as the Petsval surface, which is indicated in Fig. 11 by P.

Distortion is the variation of magnification with the distance of an object point from the axis. Positive, or "barrel" distortion, exists if the magnification decreases with increasing distance of an object point from the axis. Negative, or "pincushion," distortion exists if the magnification increases with increasing distance of an object point from the axis. Distortion increases as the cube of the distance of an object point from the axis.

Chromatic aberration is due to the fact that the focal length of a lens depends upon its index of refraction and thus upon the wavelength of the light used. As a consequence

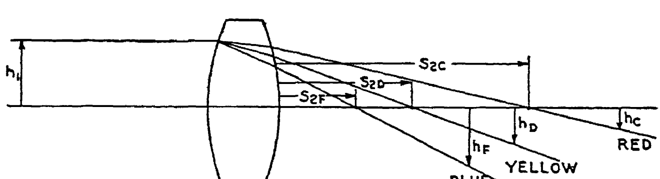


FIG. 12. Longitudinal and Lateral Chromatic Aberration

a single lens forms a series of colored images of different magnifications at different distances from the lens (Fig. 12). The variation of image distance with wavelength is known as *longitudinal chromatic aberration.* The difference $S_{2C} - S_{2F}$ is taken as a measure of the longitudinal aberration, and for a thin lens it is given by

$$s_{2C} - s_{2F} = \frac{1}{f_{2D}V\left(\dfrac{1}{f_{2D}} + \dfrac{1}{S_{1D}}\right)^2} = \frac{S_{2D}^2}{f_{2D}V} \qquad (69)$$

where V is given by eq. (6). The variation of image size with wavelength of refraction is known as *lateral chromatic aberration* and is measured by $h_C - h_F$.

STOPS. Any obstacle which limits the rays that can be transmitted through an optical system is a stop. Stops serve (1) to limit the aperture of any bundle of rays, and thus the amount of light, that reaches any image point, and (2) to limit the extent of the object which is imaged, or the field of view. The stop that limits the diameter of the bundle of rays that can enter the optical system is called the aperture stop or iris. The stop that limits the field of view is known as the field stop. In Fig. 13, NP is the image of AS formed

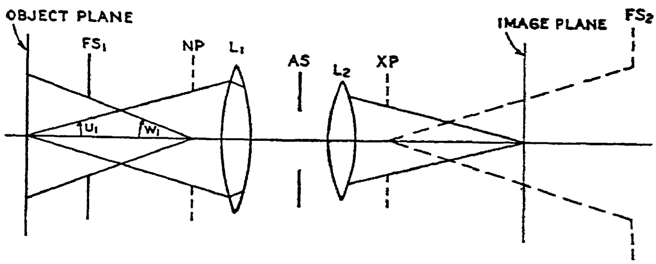


Fig. 13. Aperture Stop, AS, and Field Stop, FS_1, in an Optical System

by that part of the optical system (lens L_1) situated between AS and the object. Since NP limits the angular aperture U_1 of the beam, AS is the aperture stop. NP is known as the entrance pupil. If an actual stop were located at NP instead of AS, then the iris and entrance pupil would coincide. The image XP of AS formed by that part of the optical system (lens L_2) situated between AS and the image is known as the exit pupil. The angle subtended by the object at the center of the entrance pupil is known as the *angular field of view* of the object. In Fig. 13 the angular field of view of the object is limited to $2W_1$ by the aperture stop FS_1. If it is desirable to have a sharp boundary of the field of view it is necessary to have FS_1 in the plane of the object.

Stops play a very important role in eliminating those rays that would produce excessive aberrations.

DEPTH OF FOCUS. The maximum total distance, ΔS_2, that a viewing screen may be moved along the axis of an optical system to produce a circle of confusion of prescribed radius, r, is called the depth of focus (Fig. 14a). It is given by

$$\Delta S_2 = 2r\frac{S_2}{a_2} = 2r \cot U_2 \qquad (70)$$

where S_2 is the image distance, at best focus, measured from the exit pupil of radius a_2.

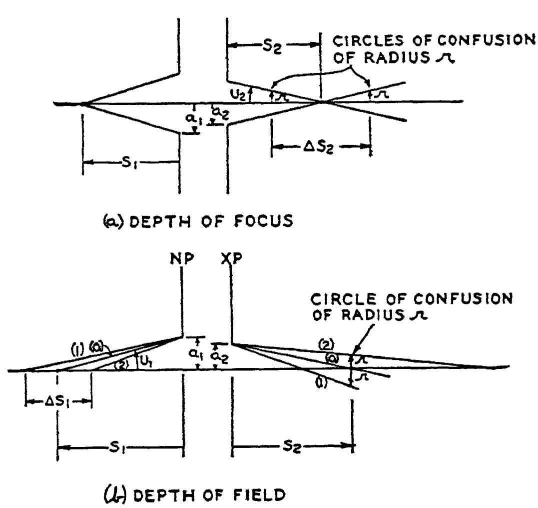


Fig. 14. Difference between Depth of Field and Depth of Focus

DEPTH OF FIELD. The maximum total distance, ΔS_1, that an object may be moved along the axis to produce, at a fixed image distance, a circle of confusion of prescribed radius, r, is called the depth of field (Fig. 14b). It is given by

$$\Delta S_1 = \frac{rS_1}{ma_1 - r} + \frac{rS_1}{ma_1 + r} \qquad (71)$$

where m is the magnification and S_1 is the object distance, at best focus, measured from the entrance pupil of radius a_1. If $r \ll ma_1$, then

$$\Delta S_1 = 2\frac{r}{m}\frac{S_1}{a_1} = 2\frac{r}{m}\cot U_1 \tag{72}$$

3. PHOTOMETRY

The problem in photometry is to obtain a quantitative evaluation of radiant flux with respect to its capacity to produce the *sensation of brightness*. The sensation of brightness evoked by a given amount of radiant energy is different for different individuals and is different for the same individual under different conditions of observation. Equal amounts of radiant energy per unit wavelength interval throughout the visible spectrum do not produce visual sensations of equal brightness. A luminosity curve shows as ordinates the relative effectiveness of various wavelengths to evoke, for a particular observer, visual sensations of equal brightness. Figure 15 shows two luminosity curves applying to

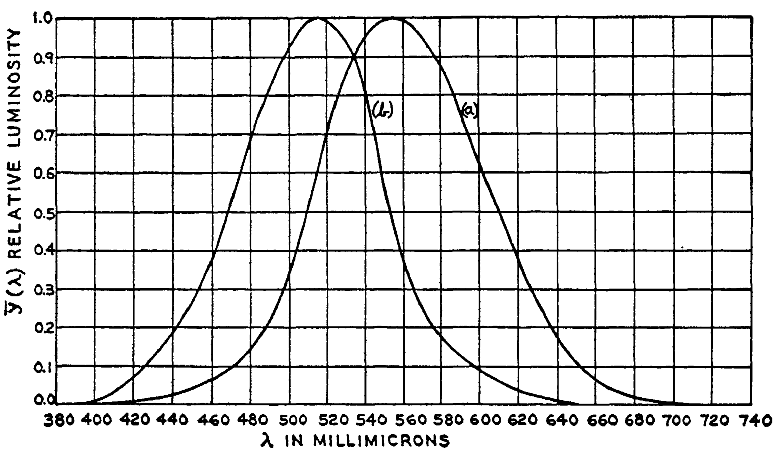


FIG. 15. Curve (*a*), Standard Luminosity Function, Applying to Normal Vision with Good Lighting Conditions; Curve (*b*), Luminosity Function Applying to Vision at Very Low Light Levels

photopic or normal vision (at high light levels) and to scotopic vision (at very low light levels). These are average luminosity curves obtained for a number of observers. The use of many luminosity curves would, obviously, lead to endless confusion and ambiguity in photometric measurements and specification. To avoid this, the International Commission on Illumination adopted Fig. 15*a* as the *standard luminosity curve*. It is important to realize that the standard luminosity curve is essentially an arbitrarily assumed standard for purposes of standardization and specification of photometric data and is not necessarily the retinal response of any individual. Table 3 gives the standard luminosity function, $\bar{y}(\lambda)$ of Fig. 15*a*, at 10-millimicron intervals.

Table 3. Standard Luminosity Function

λ in mμ	$\bar{y}(\lambda)$	λ in mμ	$\bar{y}(\lambda)$	λ in mμ	$\bar{y}(\lambda)$	λ in mμ	$\bar{y}(\lambda)$
380	0.0000	480	0.139	580	0.870	680	0.0170
390	.0001	490	.208	590	.757	690	.0082
400	.0004	500	.323	600	.631	700	.0041
410	.0012	510	.503	610	.503	710	.0021
420	.0040	520	.710	620	.381	720	.0010
430	.0116	530	.862	630	.265	730	.0005
440	.0230	540	.954	640	.175	740	.0003
450	.0380	550	.995	650	.107	750	.0001
460	.0600	560	.995	660	.061	760	.0001
470	.0910	570	.952	670	.032		

Table 4 gives the names, symbols, and basic mks units of radiometry and photometry as recommended by the committee on colorimetry of the Optical Society of America. The term "luminance" in Table 5 replaces the older term "brightness" which led to confusion between the objective concept of brightness as a measurable quantity and the subjective concept of brightness which refers to the sensation in the consciousness of the human observer. It is recommended that the term "brightness" be used only in the latter sense.

Table 4

Radiometry			Photometry		
Name	Symbol	Unit MKS	Name	Symbol	Unit MKS
Radiant energy......	U	Joule	Luminous energy.....	Q	Talbot
Radiant flux........	P	Watt	Luminous flux	F	Lumen
Radiant emittance...	W	Watt/m^2	Luminous emittance..	L	Lumen/m^2
Radiant intensity....	J	Watt/ω (steradian)	Luminous intensity...	I	Lumen/ω (candle)
Radiance...........	N	Watt/ωxm^2	Luminance..........	B	Lumen/ωxm^2 (candle/m^2)
Irradiance..........	H	Watt/m^2			
			Illuminance..........	E	Lumen/m^2 (lux)

The ratio of any photometric quantity to the corresponding radiometric quantity in Table 5 is equal to the absolute luminosity or *luminous efficiency* (generally expressed as lumens per watt) of the radiant energy. Thus 1 watt of monochromatic radiant flux of wavelength 555 mμ (corresponding to the peak of the standard luminosity curve) is equivalent to 685 lumens. This efficiency of 685 lumens per watt is based on the fact that the new proposed international photometric standard of a black body at the temperature (2043.8 deg K) of freezing platinum shall have a luminance of 6×10^5 candles per m^2 (60 candles per cm^2).

Table 5. Conversion Factors for Units of Illuminance

to Obtain ↓	Multiply Number of → by	Lux	Foot-candle	Phot	Milliphot
Lumen/m^2	Lux......................	1	10.76	10,000	10
Lumen/ft^2	Foot-candle..............	0.0929	1	929	0.929
Lumen/cm^2	Phot.....................	0.0001	0.001076	1	0.001
	Milliphot................	.1	1.076	1,000	1

If the radiant flux is monochromatic its luminous efficiency is simply $685\bar{y}(\lambda)$. If the radiant flux is not monochromatic but consists of a continuous spectrum, then if $P(\lambda)$ is the radiant flux per unit wavelength (watts per millimicron) the luminous efficiency is

$$K = 685 \frac{\int_0^\infty \bar{y}(\lambda) P(\lambda)\, d\lambda}{\int_0^\infty P(\lambda)\, d\lambda} \tag{73}$$

Luminous efficiency should not be confused with the efficiency of a practical light source, which is the ratio of the total luminous flux to the total power input. The efficiency of a source of light is less than K since generally a fraction of the total power input is not converted into radiant flux.

LUMINOUS INTENSITY OR CANDLEPOWER OF A SOURCE. By a source will be understood (*a*) any self-luminous object, such as an incandescent body, or (*b*) any illuminated object which so completely diffuses (either by reflection or transmission) the incident light that it acts as a source.

The intensity of a source is defined as

$$I = \frac{dF}{d\omega} \tag{74}$$

and is measured in lumens per steradian, or candles. In Fig. 16, let the small plane source of area A_0 emit P watts or $F = 685\int_0^\infty \bar{y}(\lambda)P(\lambda)\,d\lambda$ lumens. The two small receivers of areas A_1 and A_2 subtend the same solid angle $\Delta\omega$ at the (say, center of) source and are located at the same distance D from the source. The number of lumens ΔF_1 and ΔF_2 contained in the same solid angle $\Delta\omega$ will, in general be different. Thus the intensity normal to the source, $I_0 = \frac{\Delta F_1}{\Delta\omega}$, is not equal to the intensity, $I_\alpha = \frac{\Delta F_2}{\Delta\omega}$, measured along a direction at the angle α with the normal. In specifying the intensity or candlepower of a source, it is therefore necessary to state the direction in which the intensity was measured or give a candlepower distribution curve. The directional characteristic of many extended sources follow *Lambert's law*, which states that

$$I_\alpha = I_0 \cos\alpha \tag{75}$$

With a uniform point or spherical source, the intensity is independent of direction and is $F/4\pi$. Thus a uniform point or spherical source of an intensity of 1 candle emits 4π lumens.

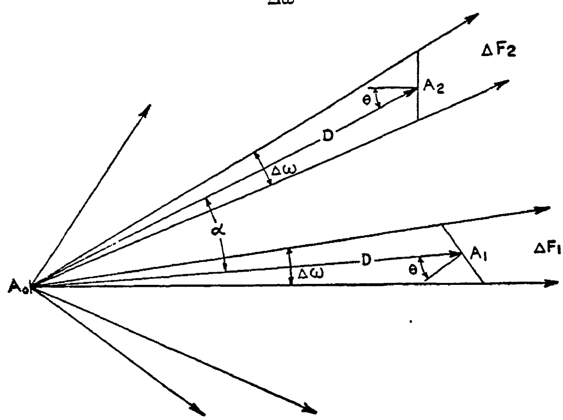


FIG. 16. Diagram Illustrating Some Photometric Concepts

ILLUMINANCE. A surface of area dA, placed in a field of flux, is illuminated with the illuminance

$$E = \frac{dF}{dA} \text{ (lumens per } m^2 \text{ in mks units)} \tag{76}$$

where dF is the flux incident on the surface. The illuminance of receiver A_1 (Fig. 16) is

$$E_1 = \frac{\Delta F_1}{A_1} = \frac{I_0 \Delta\omega}{A_1} = \frac{I_0}{A_1}\frac{(A_1 \cos\theta)}{D^2} = \frac{I_0}{D^2}\cos\theta \tag{77a}$$

The illuminance of receiver A_2 is

$$E_2 = \frac{I_\alpha}{D^2}\cos\theta = \frac{I_0}{D^2}\cos\alpha\cos\theta \tag{77b}$$

In normal incidence

$$E_1 = \frac{I_0}{D^2}, \qquad E_2 = \frac{I_\alpha}{D^2} \tag{78}$$

which is the well-known inverse-square law. The inverse-square law, which is the basis of most visual photometers, applies to large extended sources, if D is considerably larger than the extension of the source. Thus (78) holds to within about 1 per cent if D is at least 5 times the greatest linear dimension of the source.

LUMINANCE. The luminance, B, of the surface A_0 (Fig. 16) in any direction is the ratio of the intensity I_α in that direction to the area of the projection of A_0 on a plane perpendicular to the direction, i.e.,

$$B = \frac{I_\alpha}{A_0 \cos\alpha} \tag{79}$$

Luminance is measured in candles per square meter in mks units. If Lambert's law is followed, then

$$B = \frac{I_0 \cos\alpha}{A_0 \cos\alpha} = \frac{I_0}{A_0} = B_0 \tag{80}$$

and the luminance of a surface is independent of α. The *brightness sensation* when observing a surface, whether self-luminous (as the luminescent screen of a cathode-ray tube) or diffusely transmitting or reflecting (as a television or movie projection screen), depends upon the *luminance* of the surface. Hence if the surface obeys Lambert's law, its luminance is the same in all directions, and it will appear equally bright from whatever angle it is viewed.

LUMINOUS EMITTANCE. The luminous emittance, L, of a surface is the total luminous flux emitted per unit of area, or

$$L = \frac{F}{A} \tag{81}$$

Luminous emittance is measure in lumens per square meter in mks units.

The luminous emittance of a surface *obeying Lambert's law* is found to be

$$L = \pi B \tag{82}$$

Hence a perfectly diffusing (emitting, transmitting, or reflecting) surface, whose luminance is B candles per square meter has a luminous emittance of πB lumens per square meter. Or a perfectly diffusing surface of luminance B and area A emits

$$F = \pi BA \text{ lumens} \tag{83}$$

If the surface radiates on both sides with luminance B, then

$$F = 2\pi BA \tag{84}$$

PHOTOMETRIC UNITS. Equation (82) is the basis for another unit of luminance called meter-lambert. By definition 1 meter-lambert is the luminance of a perfectly diffusing surface emitting, reflecting, or transmitting, 1 lumen per m². Thus 1 meter-lambert equals $1/\pi$ candle /m². This unit (meter-lambert) is very convenient when dealing with non-self-luminous surfaces such as perfectly diffusing (transmitting or reflecting) surfaces. If these surfaces do not absorb any light, the number of lumens incident on them is equal to the number of lumens transmitted or reflected. Hence the illuminance in lumens per square meter, the luminous emittance in lumens per square meter, and the luminance in meter-lamberts are all numerically equal. If these surfaces do absorb some light, the luminance in meter-lamberts is equal to the illuminance in lumens per square meter multiplied by the fraction of the incident light that is transmitted or reflected. The luminous emittance in lumens per square meter is still equal to the luminance in meter-lamberts. If the surface does not obey Lambert's law its luminance depends upon the angle of observation and the advantage of this unit disappears. The luminance of such a surface in a particular direction in meter-lamberts may then be interpreted as the number of lumens per square meter that a perfectly diffusing surface of the same luminance would radiate.

Besides the mks units given above, there are many others in widespread use. They differ from the mks units only in being based on different units of area. Tables 5 and 6 give the names and the relative magnitudes of the more common units of illuminance and luminance.

Table 6. Conversion Factors for Units of Luminance

Multiply Number of → by to Obtain ↓	Candle per cm²	Candle per in.²	Candle per ft²	Candle per m²	Lambert	Milli-lambert	Foot-lambert (equivalent foot-candle)	Meter-lambert
Candle per cm² (Stilb)	1	0.1550	0.0010764	10^{-4}	0.3183	0.0003183	0.0003426	0.00003183
Candle per in.²	6.452	1	0.006944	6.452×10^{-4}	2.054	0.002054	0.00221	0.0002054
Candle per ft²	929	144	1	0.0929	295.7	0.2957	0.3183	0.02957
Candle per m²	10,000	1,550	10.764	1	3,183	3.183	3.426	0.3183
Lambert (cm-lambert)	3.142	0.4869	3.382×10^{-3}	3.142×10^{-4}	1	0.001	0.001076	10^{-4}
Millilambert.........	3,142	486.9	3.382	0.3142	1,000	1	1.0764	0.1
Foot-lambert.........	2,919	452.4	3.142	0.2919	929	0.929	1	0.0929
Meter-lambert........	31,420	4,869	33.82	3.142	10^4	10	10.76	1

4. LIGHT MEASUREMENT

The methods of light measurement may be divided into two classes; visual photometry and physical photometry.

In visual photometry the human eye is the detector. Although the human eye is incapable of measuring, it is capable of fairly accurately judging the equality of luminances of adjacent areas. In a visual photometer, two adjacent areas of a screen are illuminated by a calibrated source and an unknown source. The observer adjusts the illuminance on the half-field produced by the calibrated source (by varying the distance between source and screen) until he judges the two half-fields to be equally "bright." The better visual photometers (such as the "Macbeth Illuminometer") use a Lummer Brodhun cube to split the field. Relatively accurate measurements may be made with visual photometers only if the spectral distribution of the calibrated and unknown sources are approximately the same. Although a series of filters may relieve this situation, measurements upon sources of different colors (heterchromatic photometry) are generally subject to great errors unless a flicker photometer is used.

In *physical photometry*, the detector is generally a photovoltaic or photoemissive cell. A cell will give true photometric values regardless of the color of the light if corrected with a suitable filter, so that its sensitivity throughout the spectrum is proportional to the luminosity curve of Fig. 15*a*. The most important error in physical photometry is generally due to the difference between the spectral response of the cell and the standard luminosity curve. Other errors arise from the directional and temperature characteristics of cells. Most commercial light measurements are now made using a photovoltaic cell with an "eye"-corrected filter.

5. PHOTOMETRIC RELATIONS IN NON-VISUAL OPTICAL SYSTEMS

One of the important performance characteristics of an optical system is its light-gathering power. In Fig. 17 the small object of area A_1 radiates in all directions, but the

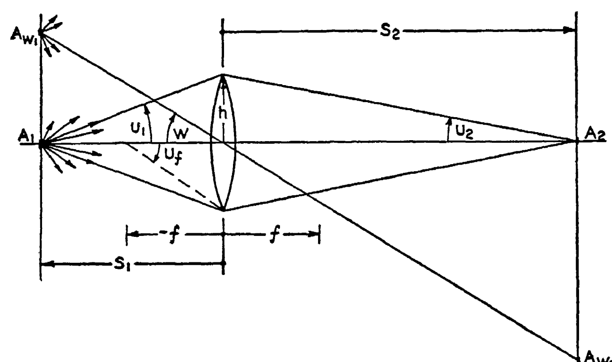


FIG. 17. Diagram Illustrating Some Photometric Relations in Optical Systems

lens accepts and focuses only a fraction of the flux on the image of Area A_2. If A_1 emits according to Lambert's law, then the fraction of the flux that the lens accepts is

$$e = \sin^2 U_1 \tag{85}$$

where U_1 is the angle subtended by the radius of the entrance pupil at A_1. The quantity e might be called the (geometric) efficiency of the optical system. If B_1 is the luminance of A_1, then the total flux, F, that A_1 emits is $\pi B_1 A_1$ and the flux accepted by the lens is

$$F_1 = \pi B_1 A_1 \sin^2 U_1 \tag{86}$$

If the lens has no losses, F_1 is the flux that will reach the image. If there are losses in the optical system, let k be the fraction of the incident light that it transmits, and

$$F_2 = kF_1 = \pi B_1 A_1 k \sin^2 U_1 \tag{87}$$

The quantity $k \sin^2 U_1$ might be called the effective efficiency of the optical system.

The illuminance of the image is

$$E_2 = \frac{F_2}{A_2} = \frac{\pi B_1 k \sin^2 U_1}{m^2} = \pi B_1 k \sin^2 U_2 \tag{88}$$

This equation is the basic relation giving the flux density on a small area (A_2 in Fig. 7) of the image, located on the axis of an aplanatic optical system, produced by an extended source emitting according to Lambert's law.

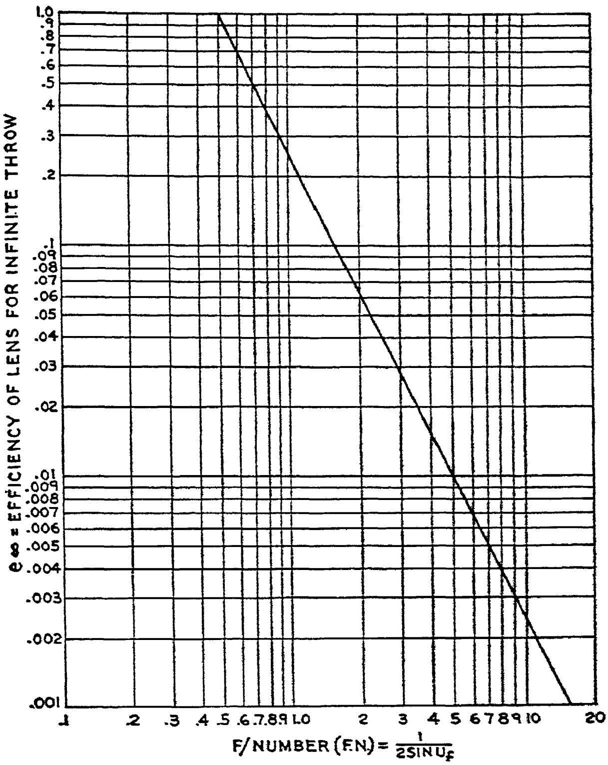


FIG. 18. Efficiency of a Lens Used for Projection at Infinite Throw as a Function of F/Number

Most lenses are rated (as to their light-gathering power) by

$$f/\text{number} = fn = \frac{f}{2h} = \frac{1}{2 \tan U_f} \tag{89}$$

and sometimes by

$$\text{Relative aperture} = RA = \frac{2h}{f} \tag{90}$$

Microscope objectives are rated by

$$\text{Numerical aperture} = NA = N_1 \sin U_1 \tag{91}$$

If the entrance and exit pupils coincide with the principal planes of the optical system, eq. (88) may be written

$$E_2 = \pi B_1 k \frac{h^2}{h^2 + S_1^2} \frac{S_1^2}{S_2^2} = \frac{\pi B_1}{4} k \frac{(RA)^2}{(m^2/4)(RA)^2 + (m-1)^2} \tag{92}$$

For camera and telescope objectives, the object distance, S_1, is usually very large and the image distance, S_2, is approximately equal to the focal length (or $m \cong 0$) so

$$E_2 = \frac{\pi B_1}{4} k(RA)^2 = \frac{\pi B_1}{4} \frac{k}{(fn)^2} \tag{93}$$

If the relative aperture is small or f/number large, then for any magnification

$$E_2 = \frac{\pi B_1}{4} k \frac{(RA)^2}{(m-1)^2} = \frac{\pi B_1}{4} \frac{1}{(fn)^2(m-1)^2} \tag{94}$$

For high-efficiency projection systems, it is preferable to define f/number as

$$FN = \frac{1}{2 \sin U_f} \tag{95}$$

instead of eq. (89). For values of U_f less than about 10°, $\sin U_f = \tan U_f$ and the two definitions agree, but they disagree for larger values of U_f. The efficiency of a projection lens for infinite throw is thus

$$e_\infty = \sin^2 U_f = \frac{0.25}{(FN)^2} \tag{96}$$

For a lens immersed in air, the smallest FN possible is 0.5, since then the efficiency is unity, and all the light emitted is accepted by the lens. For a lens immersed in a medium of index N_1 the efficiency is $e_\infty = N_1^2 \sin^2 U_f$ and the FN is $\frac{1}{2\,N_1 \sin U_f}$. Figure 18 shows the efficiency e_∞ of a lens as a function of f/number. It is seen that the efficiency of most lenses is very low.

The efficiency of a given optical system decreases when the magnification or throw decreases. The efficiency at a magnification m is

$$e_m = \frac{h^2}{h^2 + S_1^2} = \frac{e_\infty}{1 + (1 - e_\infty)\left(\frac{1 - 2m}{m^2}\right)} \tag{97}$$

Thus an ordinary lens with an FN of 2 has an e_∞ of 6.25 per cent and an e_m of about 4.6 per cent when used at a magnification of -6. Similarly a lens with an e_∞ of 25 per cent ($FN = 1$) will have an efficiency of 19.6 per cent at a magnification of -6.

The illuminance of an extra-axial area (Aw_2 Fig. 17) produced with a lens of small aperture is approximately given by

$$E_{2w} = \frac{\pi}{4} Bk \cos^4 W \frac{(\text{R.A.})^2}{4(m - 1)^2} \tag{98}$$

where w is the field angle (see Fig. 17).

6. REFLECTIVE OPTICAL SYSTEM FOR TELEVISION PROJECTION

Two important requirements of an optical system for television projection are: (1) that it be capable of focusing a large field (large tube face) and (2) that it have high efficiency. The need for a large field is that, owing to current saturation and heating of the luminescent screen, the light output from a cathode-ray tube increases with the size of the tube, for a given beam power input.

The most efficient optical systems capable of focusing considerable fields are, in general, of the reflective type. One of the simplest (and best) of these consists of a spherical mirror and an aspherical aberration-correcting lens located at the center of curvature of the mirror (Fig. 19).

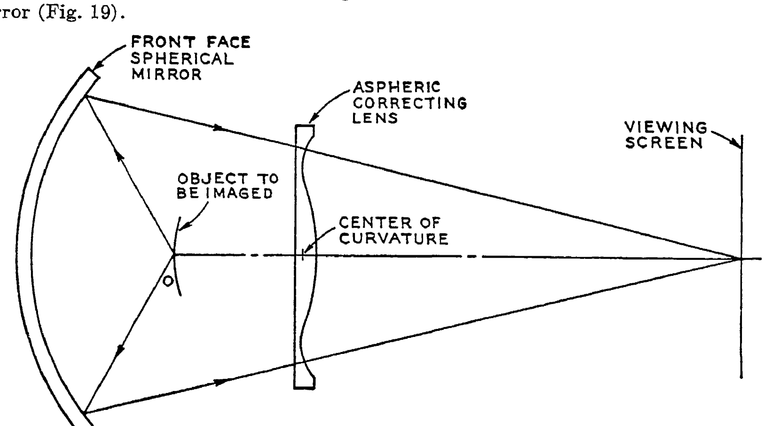


FIG. 19. Reflective Optical System for Television Projection

The aberrations of any optical system with a high geometrical efficiency can be sufficiently corrected for only one position of object and image. As a result a given system is good for only a small range of magnification, and different systems have to be designed for different magnifications or throws. A highly efficient system has, because of the large convergence angles, relatively shallow depth of focus and depth of field; see eqs. (70) and (71).

FOCUSING WITH MIRROR AND CORRECTING LENS. Because of the symmetry of the sphere, an optical system consisting of a spherical mirror and a small aperture located at the center of curvature of the sphere suffers from only two aberrations: spherical aberration which is uniform all over the field, and curvature of the field. With large apertures, extra-axial aberrations of higher order enter the field, but they are not too large (see Fig. 20).

The purpose of the correcting lens is to correct for the spherical aberration of the mirror without introducing any serious aberrations of itself. This is accomplished by making

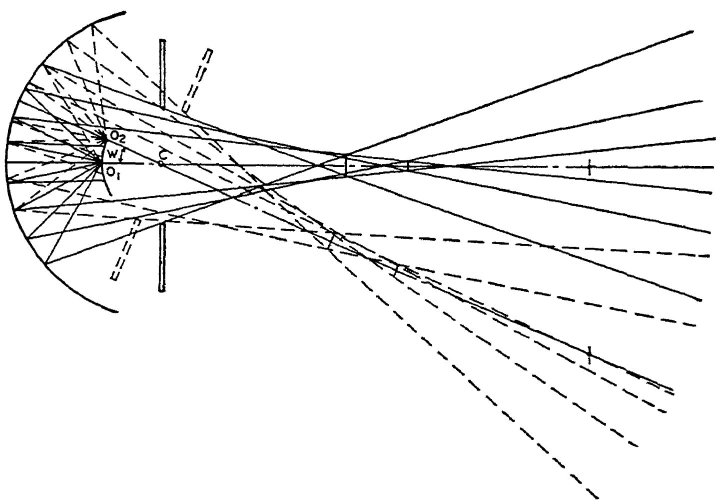


FIG. 20. Image Properties of System Consisting of Spherical Mirror and Aperture Located at Center of Curvature

the lens as weak as possible and locating it in the plane of the aperture at the center of curvature. In this way the symmetry property of the spherical mirror is least disturbed. The curvature of the field is not corrected as it is actually used to good advantage in cathode-ray-tube projection.

Equation (20) may be written

$$\frac{1}{P_1} + \frac{1}{P_2} = -\frac{2\cos U_0}{R} = \frac{1}{f_{U_0}} \tag{99}$$

where f_{U_0} is the focal length of a zone of the mirror of aperture sin U_0. The zonal focal length f_{U_0} thus increases with the aperture of the zone.

The spherical aberration of the mirror may be interpreted as focusing by means of zones, each zone having a different focal length. The correcting lens has to be such that each zone of the lens has a different focal length, compensating for the various focal lengths of the mirror and resulting in a focusing system with all zones of the same focal length.

The shape of the correcting lens will thus depend upon the zonal focal length of the mirror one chooses as the focal length of the optical system (mirror plus correcting lens). Since theoretically there are an infinite number of zones on the mirror, there are theoretically an infinite number of correcting lens shapes that will produce a system in which all zones have the same focal length.

Since the mirror with an aperture at the center of curvature has no extra-axial or chromatic aberrations, such aberrations are caused by the correcting lens itself, i.e., by the power or slopes on the correcting lens. From the standpoint of these aberrations, therefore, that shape should be chosen whose maximum slope is the least. Thus if the paraxial (central) focal length of the mirror is chosen as that of the system, then the central focal length of the correcting lens is infinite and the shape of the curve is concave. If a zonal focal length of the mirror is chosen as that of the system there will be a zonal focal length of the correcting lens which is infinite and the shape of the curve is convex at the center and concave past this zone. If a peripheral focal length is chosen, the required correcting lens is convex. The maximum slope is least for a convex-flat-concave curve.

The shape of the correcting lens must be such that all rays emanating from an object point O_1, and reflected by the mirror, shall meet at the image point O_2 located at a distance S from the correcting lens. Figure 21 shows three rays emanating from O_1 and

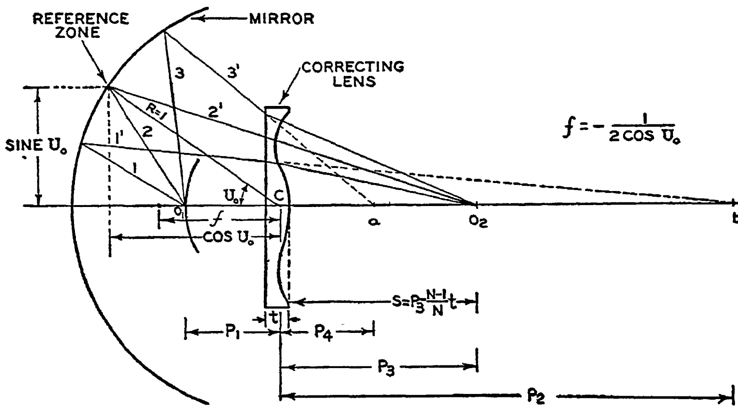


FIG. 21. Diagram Illustrating the Effect of the Correcting Lens

striking the mirror at different apertures. Without the presence of the correcting lens, rays 1, 2, 3 would intersect the axis at distances P_2, P_3, and P_4 from the center of curvature. The slopes on the correcting lens have to be such (approximately as shown on Fig. 21) that all three rays intersect at O_2; hence, the correcting lens has a flat zone at the point where ray 2′ passes, negative slope where ray 1′ passes, and positive slope where ray 3′ passes.

From the point of view of spherical aberration, if the zone where ray 2 strikes the mirror is taken as a reference, then the mirror has negative spherical aberration for smaller apertures and thus requires a positive lens for correction, and positive spherical aberration for larger apertures and thus requires a negative lens.

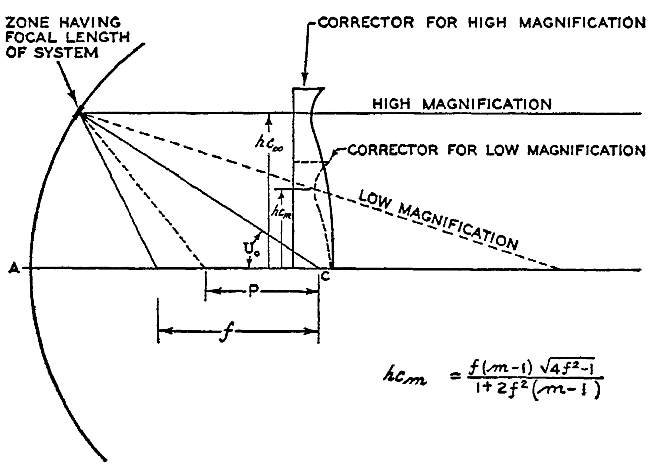


FIG. 22. Diagram Showing Why the Aperture of the Correcting Lens Depends upon Magnification

The shape and size of the correcting lens depend upon the throw or magnification for which the system is to be used. For a given focal length and relative aperture the correcting lens aperture decreases as the magnification decreases (see Fig. 22). That this must be so may be surmised from the fact that for unity magnification the lens aperture is zero, since object and image coincide at the center of curvature. Figure 23 shows the variation of correcting lens semiaperture and mirror semiaperture, with magnification, for a system of given focal length and efficiency. All distances in Fig. 23 are measured in terms of the radius of curvature of the mirror; i.e., the radius of curvature is taken as the unit of length.

The focal length of the complete optical system depends upon the shape of the correcting lens. In general, the higher the efficiency for which a system is designed the greater the focal length (in terms of R). Thus for an extremely inefficient system (requiring only a small aperture at the center of curvature and no correcting lens) the focal length would

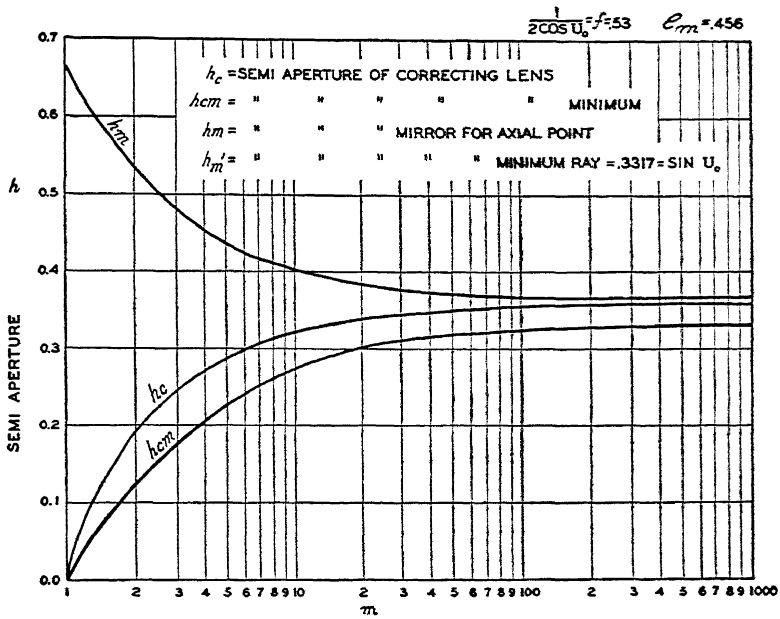


FIG. 23. Variation of Semi-aperture of Mirror and Correcting Lens with Magnification

be $0.5R$, whereas a reasonably good design of a system with an efficiency of about 50 per cent (requiring a correcting lens) will have a focal length of about $0.53R$.

Relations between the throw S, magnification m, object distance P_1, and the focal length f are:

$$S = mP_1 \tag{100}$$

$$P_1 = f\frac{(m+1)}{m} \tag{101}$$

$$S = -f(m+1) \tag{102}$$

Thus for a magnification of -6 (inverted image) and a focal length of $0.53R$ the throw is $2.65R$. In the reflective optical systems contemplated for home television projection receivers $R = 13.7$ in., and so for a magnification of -6 the throw would be 2.65×13.7 in. $= 36.3$ in.

TUBE FACE. Before striking the spherical mirror, the light emitted by the luminescent screen of the cathode-ray tube first passes through a thickness (generally about $1/8$ in.) of glass constituting the tube face. The tube face should preferably consist of two concentric surfaces, and thus it acts as a weak lens. The radius of curvature of the outer surface (for a thin tube face) is approximately equal to the focal length of the system. The lens action of the tube face changes the magnification of the system to

$$M = mm_t = \frac{S}{P_1}\frac{1}{1 - \dfrac{N-1}{N}\dfrac{t}{R_1}} \tag{103}$$

where N, t, and R_1 are the index of refraction, thickness, and radius of curvature of the outer surface of the tube face. However, the largest effect of the tube face is caused by images cannot be seen. The blind spot is roughly elliptical, the vertical dimension being longer. Its size varies greatly among individuals, with limits for horizontal dimensions of 3 to 8 degrees, averaging about 6 degrees. The nearest edge is located about 12 degrees on the nasal side of each retina.

Contrary to other sense organs, the entire nerve system leading to the cortical regions of the brain is found essentially in the retina itself. Here some of the functions may occur which usually take place in the cortex, for example the color-analyzing processes (Polyak).

THE REFRACTIVE MEDIA. The transparent intraocular fluids in all parts of the eye are derived from the blood and physiologically are essentially the same. The fluid in

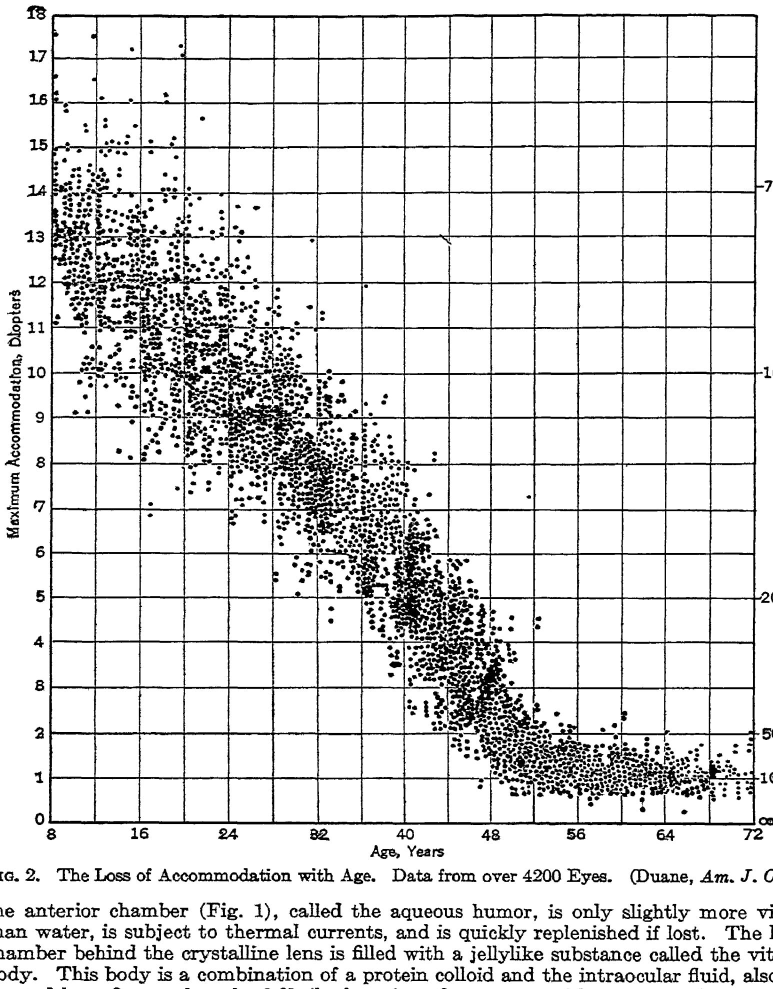


FIG. 2. The Loss of Accommodation with Age. Data from over 4200 Eyes. (Duane, *Am. J. Ophth.*)

the anterior chamber (Fig. 1), called the aqueous humor, is only slightly more viscous than water, is subject to thermal currents, and is quickly replenished if lost. The larger chamber behind the crystalline lens is filled with a jellylike substance called the vitreous body. This body is a combination of a protein colloid and the intraocular fluid, also permeated by a fine meshwork of fibrils that gives the mass a stable anatomical structure.

The crystalline lens is a transparent body having the shape of a biconvex lens, and it serves the same function as a lens. It consists of a non-homogeneous elastic substance made up of a number of layers or laminae, each with increasing density and increasing index of refraction toward the core at the center. By the action of the ciliary body, through a complex process, the lens may become more convex and in so doing serves the function of changing the position of the focal point of the light entering the eye. In this way, the eye can accommodate itself from distinct vision of distant objects to that of nearer objects, and vice versa. In this act of accommodation primarily the radius of the front surface decreases and the thickness of the lens increases.

With increasing age the average density of the lens increases; the lens becomes harder and less elastic and hence its accommodative function is less and less effective. The loss of accommodation with age is illustrated in Fig. 2 (Duane), where the nearest distance from the eye at which a test object remains clear is plotted against age. It is clear that, at about the age of 60 years, little or no accommodation remains, and the eye is then said to be completely *presbyopic*.

THE PUPIL. The pupil acts as the stop or aperture of the eye. It is in a constant state of activity and is subject to a number of reflexes. Although the pupil will contract when the iris is stimulated directly by light, normally it contracts through a light stimulation of the retina, the extent depending on the adaptation of the eye. The pupillary response to light falling on the central part of the retina is much greater than for that falling on the peripheral parts of the retina. The threshold for pupillary response of the central region of the retina, when dark adapted, corresponds quite well to the threshold of cone vision. Data showing the change in the diameter of the pupil in response to the luminosity of an extended surface are illustrated in Fig. 3 (Reeves). Theoretically the amount of light entering the eye would be approximately proportional to the area of the pupil (diameter squared), but it is evident from the figure that, whereas the illumination change is more than several million times, the amount of light entering the eye changes only about 15 fold. The contraction of the pupil cannot, therefore, even approximately compensate for increasing illumination. The rate of contraction of the pupil from darkness to light is much greater than the rate of dilation resulting from light to darkness. The response of the pupil to colored light is greater for yellow than for red or blue.

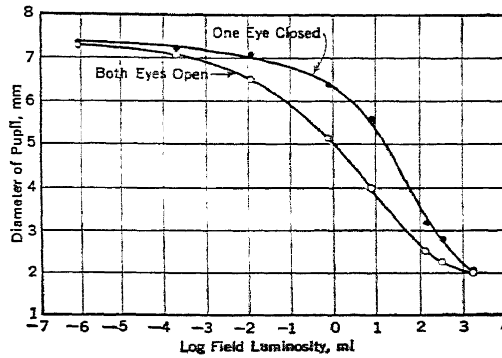


FIG. 3. The Relation between Pupillary Size and the Luminosity of the Visual Field. (Data of Reeves, *J. Optical Soc. Am.*)

In this connection, it has been shown (Stiles and Crawford) that the intensity of the light falling upon the retina is not directly proportional to the pupil area. Rays entering the pupil near the edge are much less effective in producing a sensation than those entering the center of the pupil. The relative effectiveness of the different pupillary zones is shown in Fig. 4.

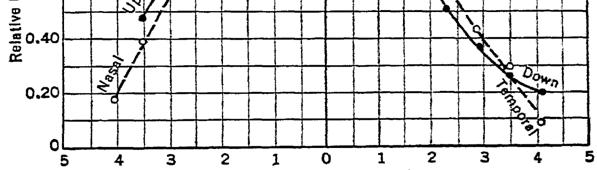


FIG. 4. Data Showing the Stiles-Crawford Effect. Light through peripheral zones of the pupil is less effective than that through the center. (From Moon, *Scientific Bases of Illuminating Engineering.*)

The pupil of the eye contracts when the eyes are converged and accommodated for near objects. This association is more closely related to convergence of the eyes than to accommodation. The reflex takes place in addition to changes in pupillary reaction due to changes in illumination. Automatic constriction of the pupil for near objects increases the depth of focus of the eye, where an increased depth is useful. Under normal conditions pupillary action takes place nearly equally in the two eyes, even if only one eye is subjected to illumination changes. The pupil is also influenced by psychic factors; a dilation is usually the rule, except when the individual is in a comatose state. There is some evidence that pupillary contraction can be conditioned.

SCHEMATIC EYE. Because of the non-homogeneity of the crystalline lens, and the general asymmetry of the optical elements, a schematic eye can be only an approximation to the living eye. Frequently, however, such a schematic eye is useful for reference and for optical problems. The data utilized by Gullstrand for the relaxed eye are given in Table 1. The greater part of the refractive power of the eye is due to the cornea (43 diopters), the lens contributing only $^1/_3$ of the total power. Since the distance of the second focal point from the second nodal point of the schematized eye is about 17 mm, an object subtending an angle of 1 minute of arc will subtend a linear distance on the retina of 5 μ (0.005 mm). In problems where the image on the retina must be considered blurred, it is convenient to know the approximate positions of the entrance and exit pupils. All optical imagery calculations can be referred to those positions if the magnification of the two stops are known. The center of the blurred image on the retina may, for practical purposes, be taken as the point of maximum light intensity.

Table 1. The Gullstrand Schematic Eye Relaxed for Distant Vision

Position of surfaces		
Cornea, anterior surface	0.0	mm
Cornea, posterior surface	0.5	
Lens, anterior surface	3.6	
Core, anterior surface	4.15	
Core, posterior surface	6.56	
Lens, posterior surface	7.2	
Radii of curvature		
Cornea, anterior surface	7.7	mm
Cornea, posterior surface	6.8	
Lens, anterior surface	10.0	
Core, anterior surface	7.91	
Core, posterior surface	−5.76	
Lens, posterior surface	−6.0	
Refractive indices		
Cornea	1.376	mm
Aqueous and vitreous humors	1.336	
Outer portion of lens	1.386	
Core	1.406	
Complete Optical System of the Eye		
Position of first principal point	1.35	mm
Second principal point	1.60	
First nodal point	7.08	
Second nodal point	7.33	
Anterior focal length	−17.05	
Posterior focal length	22.78	
Refractive power of eye	58.64	diopters
Position of entrance pupil from cornea	2.0	mm
Position of exit pupil from posterior surface of crystalline lens	3.7	
Magnification of exit to entrance pupils	0.923	

EYE MOVEMENTS. The rotary movements of each eye are controlled by six (extrinsic) muscles; the opposing external and internal recti muscles which provide movements for looking to the right and left; the superior and inferior opposing recti-muscles which provide movements for looking upwards and downwards; and the two oblique muscles which provide torsional movements about the axes of fixation, as well as movements opposing the recti in certain eye positions. The innervations to the muscles for movements of the two eyes are said to be reciprocal in that a given movement will result from the contraction of one muscle and the relaxation of its antagonist. There is, therefore, a precise coordination of the muscles that leads to very delicate and accurate movements of the eyes. In general it can be said that the ocular movements are for the purpose of directing the eyes to the object of attention and preventing *diplopia* (double vision). The movements seem, essentially, to be reflex movements following the direction of attention and accordingly have been called psycho-optical reflexes. Because the actual movements of the eyes from one point in the visual field to another appear to be approximately correct, it has been assumed that the retinal elements have "motor values," differences in which lead to correct innervations for eye movements. The reaction time between the attention and the beginning of the eye movement, though varying with circumstances, averages between 0.17 and 0.20 sec.

Owing to the constant tonus of the muscles, the eyes are in a continuous state of activity. Thus, even with constant fixation, there are occasional large, jerky movements that average 4 minutes of arc and occur at intervals of 1 to 2 sec. In between these are smaller swinging movements, and superimposed on both are very small vibratory movements.

The voluntary movements are said to be *conjugate* if in the same direction, and *disjunctive* if in the opposite direction, as when the eyes converge for a near object. The limits within which the eyes can move without head movement determine the field of fixation. Obviously, this will vary with anatomical features of the head, but it is illustrated typically in Fig. 5. The disjunctive movements are primarily concerned with convergence movements within which binocular fusion can be maintained, and this is of importance in ophthalmology since abnormalties in this function often lead to ocular discomfort. The "far point" of convergence, which may be behind the head (a divergence of the eyes), can best be measured by ophthalmic prisms, while the "near point" can be measured by bringing a small object nearer and nearer to the nose until doubling occurs. Abnormalities in convergence are frequently associated with refractive errors but may have more deep-seated innervational origins. If one eye is covered while the other fixates a given point, the covered eye may deviate from the direction of the point. This deviation, when binocular vision is prevented, is called *heterophoria*. If the eye deviates outward, inward, upward, or twists about the fixation axis, the phorias are said to be exophoria, esophoria, hyperphoria, and cyclophoria respectively. Almost everyone exhibits at least a small phoria for certain visual distances, and in fact an exophoria of about 2 to 3 degrees for near objects would be considered normal. Phorias frequently are indications that sustained efforts are being made to maintain proper convergence of the eyes in binocular vision.

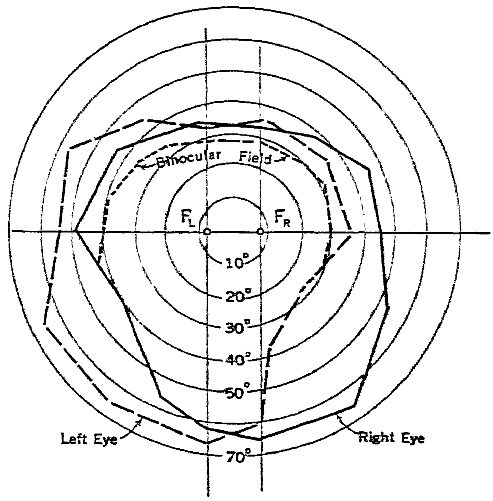


FIG. 5. An Illustration of Typical Monocular and Binocular Fields of Fixation (from Asher, *Arch. f. Ophthal.*)

In reading or similar visual tasks, the eyes make a series of short interrupted movements called saccadic movements. Even between two points, the movements do not occur exactly but may arrive by a series of small successive approximations, one eye usually leading. Eye movements are generally quicker in the horizontal than in the vertical directions, and lateral movements are faster than convergence movements. Convergence movements are more rapid than divergence movements. The speed of movement varies with the excursion, attention, and other conditions, but on the average is between 100 and 200 degrees per second.

8. THE OPTICAL CHARACTERISTICS OF THE EYE

ABERRATIONS. In addition to the usual aberrations, the image on the retina of the eye suffers also from irregular defects due to non-homogeneities in structure and the lack of symmetry in the optical media. The axis of the crystalline lens is tipped and decentered with respect to that of the cornea. The visual axis, which is determined by the fovea as well as the pupil, is usually decentered with respect to both the axes of the cornea and the lens. The spherical aberration is asymmetrical. From the central zone of the pupil to the outside zones the eye tends to be myopic, increasing at first rapidly, then reaching a maximum at 1 mm, beyond which the myopia decreases slowly. The problem of the spherical aberration of the eye is a complex one, and there are not yet enough data to understand it satisfactorily. The chromatic aberration of the eye from the C (656 mμ) to the F (487 mμ) lines of the hydrogen spectrum amounts to about 0.7 diopter. The red rays are focused to a point beyond the retina; blue rays, to a point anterior to the retina. The eye is said to be hyperopic to red, myopic to blue light. The change in chromatic foci for equal changes in wavelength is much less toward the red but increases markedly toward the blue. No entirely satisfactory explanation has been given as to why chromatic halos about images of point light sources are not seen by the eye. For emmetropes (individuals with no refractive error) the eye focuses approximately for the D (580 mμ) or yellow line of hydrogen.

Astigmatism at oblique incidence is a larger aberration. The primary (tangential) astigmatic field differs only slightly from the surface of the retina out to 20 degrees, while the secondary field rapidly increases in the myopic sense to nearly 1 $^1/_4$ diopters at a peripheral angle of 20 degrees. There is very little coma in the eye, and the curvature of the field approximates the curvature of the retina itself. The retinal image is believed to have a slight barrel distortion, though this is difficult to determine because of the subjective asymmetries of the retinas.

These aberrations in themselves, together with diffraction phenomena, would tend to reduce greatly the resolving power of the eye. However, as Gullstrand has pointed out, these aberrations together produce a complex caustic surface in the cone of light converging toward the retina, so that actually the area of maximum light concentration in the image may be small. This makes possible a resolving power whose limit is determined only by the diameter of the receptor nerve endings.

REFRACTIVE ERRORS. The eye can be defective as the result of anomalies in the refractive media or in the length of the eyeball, when it is said to be *ametropic*. The result is blurred retinal imagery, with the attendant loss in sharpness of vision. Regular ametropia is of two types—spherical and astigmatic. The first results from symmetrical abnormalities of the refractive media, or from an increased or decreased length of the eyeball. Astigmatic errors result only from non-sphericity of the surfaces of the cornea and/or of the crystalline lens. When the image focuses inside the position of the retina, the eye will be nearsighted (*myopic*), because only when the objects are brought to the eye will the images move back to the retina and appear sharply defined. In the reverse case, when the point of focus is behind the retina, the eye is said to be farsighted (*hyperopic*), because nearer objects are the more blurred. Nearsightedness can be corrected with a minus (diverging) ophthalmic lens of proper power before the eye; farsightedness with a positive (converging) lens. When astigmatism is present, the light, after entering the eye, focuses not as a point but as two lines separated in space and at right angles to each other. Images of lines in space which are parallel to that focal line nearest the retina will be seen more distinctly than those lines at right angles. For astigmatism, the corrective ophthalmic lens before the eye must have one toric surface. Astigmatism may (and usually does) occur in combination with spherical refractive errors.

Usually, when a refractive error is corrected by an ophthalmic lens, the magnification of the retinal image cannot be predicted. A spherical refractive error due to an elongation (or shortening) of the eyeball, when corrected by a lens, results in practically no magnification (or diminution) of the retinal images. If the refractive error is due to abnormalities in the refractive media, then a magnification (if the correcting lens is for farsightedness) or a diminution (if the correcting lens is for nearsightedness) will occur. The degree of magnification (or diminution) will be roughly 1 $^1/_2$ to 2 per cent per diopter power of the correcting lens. Differences in the magnification of the images of the two eyes may result in discomfort.

DEPTH OF FOCUS. As in any optical instrument, the image falling upon the retina of the eye can be slightly blurred, because of the eye's being out of focus, without an appreciable loss of perceived definition of the image. Thus, with the eye accommodated (focused) for a given distance, objects somewhat nearer and farther than this distance will appear clear. The distance between the two limits of visual distance within which the images on the retina appear not to suffer loss of definition is called the depth of focus. This fact arises, in part, because the maximum light concentration is in the center of the blur circle. Only when the intensity of this central portion is reduced as the blur circle is broadened will there be a loss in definition. The spherical and other aberrations of the eye are such that the depth of focus is greater than might be anticipated from ordinary theoretical expectations. Owing to the constriction of the pupil with accommodation and convergence, the depth of focus increases with the nearness of the fixation object. For distance vision the depth of focus, measured as the difference of the reciprocals of the nearer and farther limiting distances (in meters) for clear vision, is about 0.3 diopter, which increases to about 0.7 diopter for a visual distance of 20 cm. With the eyes relaxed for distant vision, and taking into account that the conjugate point to the retina is usually not at infinity, one could expect all objects beyond 12 ft to be clearly defined. With the eyes accommodated for the reading distance of 16 in. from the eyes, all objects from about 14 $^1/_2$ to 17 $^1/_2$ in. will appear clearly defined.

9. THE LIGHT SENSE

THRESHOLD OF LIGHT VISIBILITY. It must be clear that all determinations of visual thresholds involve problems in psychometrics (Guilford). A given weak stimulus may not always result in a definite response, and a still weaker stimulus will be responded to even less often. One determines, then, the percentage of responses for a number of exposures to the same stimulus strength. The stimulus strength is then varied in fixed steps. What percentage of responses corresponds to the threshold is a matter of definition and varies with the experimenter. It is well to know this in trying to understand and compare threshold data.

As a device for detecting radiant energy, the eye is exceedingly sensitive, though its response is confined to a narrow region of the spectrum, from about 360 $m\mu$ to 750 $m\mu$. The threshold of perception of light varies, of course, with the wavelength of the light, the size of the stimulus, and the length of exposure to the stimulus. It is also different for the cone and rod systems of the retina. Under optimum conditions, and taking the threshold as the 60 per cent response to exposures, Hecht found that with different observers the threshold varied between 2.2 and 5.7×10^{-10} erg measured at the cornea, which amounts to about 58 to 148 quanta of light energy. In this experiment a test light stimulus of wavelength 510 $m\mu$ was used, which subtended a visual angle of 10 minutes of arc, and which was exposed as a flash of 0.001-sec duration. The test light was arranged to stimulate rod vision 20 degrees temporally from the fovea, and the subject was, of course, completely dark adapted.

The relative thresholds of visibility for complete dark adaptation of the rods and cones of the fovea for different wavelengths are illustrated in Fig. 6 (Wald). Here the spectral sensitivities (reciprocal of the threshold stimulus strength) are plotted. The stimulus was a circular field subtending a visual angle of 1 degree, and this was exposed for $1/25$ sec. For the rod vision a point 8 degrees above the fovea was used. It is clear from the figure that the rods at an angle of 8 degrees into the periphery of the retina are 2.5 times more sensitive than the cones at the fovea. The maximum sensitivity (lowest threshold) for the cones occurs at 562 $m\mu$ whereas that for the rods occurs at about 505 $m\mu$. The displacement of the maxima toward the blue end of the spectrum from cone to rod vision is known as the *Purkinje* phenomenon. The fact that at the red end of the spectrum the rods and cones have essentially the same threshold sensitivities is also of special importance.

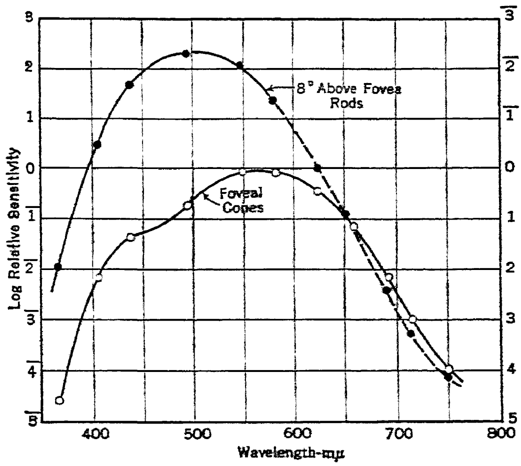


FIG. 6. Spectral Sensitivities of the Rods and Cones of the Eye, Expressed Relative to the Maximum Sensitivity of the Fovea (Wald, *Science*)

The exact relationship between the area of the light stimulus and its intensity just at the threshold of visibility is not entirely known. For the fovea, with areas less than 10 minutes of arc, and for the periphery, for areas between 2 and 7 degrees, the product of the area and intensity is approximately a constant. This fact is known as Ricco's law for the fovea and Piper's law for the periphery. These laws are wholly inadequate for wider ranges of area, and, in general, for larger areas the product of area and intensity appears to be a decreasing function of area.

In a somewhat similar manner, with a constant area, the relationship between intensity and the duration of the stimulus is not fully understood. For brief flashes of duration of less than 0.2 sec, and with small areas, the product of duration of the flash and the intensity is approximately constant (Bloch's law). However, for longer durations this constancy no longer holds, and finally the intensity alone becomes the determining criterion for threshold perception. It has also been found that the stimulation of one part of the retina by a small light source depresses the sensitivity of other parts of the retina for simultaneous but not for succeeding stimuli.

SPECTRAL LUMINOSITIES. The so-called visibility curves for daylight (*photopic*) vision and twilight (*scotopic*) vision are obtained in brightness-matching experiments. A small area illuminated by a narrow region of the daylight spectrum is presented adjacent to a standard field of constant brightness and one to which the eye is adapted. The subject adjusts the intensity of the colored light in the test area so that it is apparently equal to the brilliance of the standard field. The reciprocal of the intensity of the transmitted spectral colors for the match measures the relative brightness values of the visible spectrum. The data are obtained when the brightness of the standard field and the surroundings correspond to average daylight intensities and again to twilight visual conditions. Figure 15, p. 14-14, shows the curves which are usually found but in which the maxima are adjusted for the same height. With luminosities above about 0.01 lumen per sq ft, the photopic or cone visibility curve, which has a maximum at about 555 mμ, is found, while with luminosities below about 0.001 lumen per sq ft, the scotopic or rod visibility curve is found, with a maximum at about 507 mμ. For brightness levels between 0.01 and 0.001 lumen per sq ft there is a gradual shift of the position of the curve, this being a transition from cone to rod vision. The displacement of the maximum is again the Purkinje phenomenon. In very low illuminations the eye responds less and less to the red while still responding to the blue. No Purkinje displacement is found for foveal vision alone.

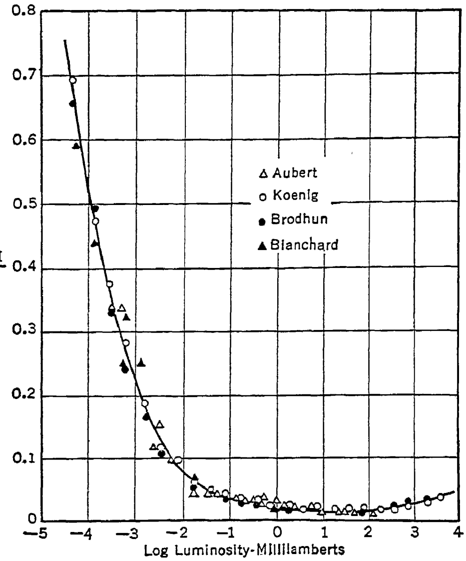


FIG. 7. Typical Data Showing the Differential Threshold of Brightness Discrimination (from Hecht, *J. Gen. Physiol.*)

DIFFERENTIAL SENSITIVITY. The smallest difference in brightness that can just be perceived (just-noticeable difference) between two areas is a measure of the differential threshold of visibility from which one arrives at a value of the differential sensitivity, or contrast sensitivity. The differential threshold varies greatly with the illumination, the wavelength, the size and separation of the test areas, and the luminosity of the surroundings. The differential threshold is expressed as the ratio $\Delta I/I$, where I is the luminosity of the standard area and the adapting field, and ΔI is the difference in luminosity of the test area and the standard area for the just-noticeable difference in brightness. The contrast sensitivity is the reciprocal of this quantity. Typical differential threshold results are shown in Fig. 7, showing data recomputed by Hecht. Because of the wide range in luminosities necessary, the abscissas are plotted on a logarithmic scale. In the photopic

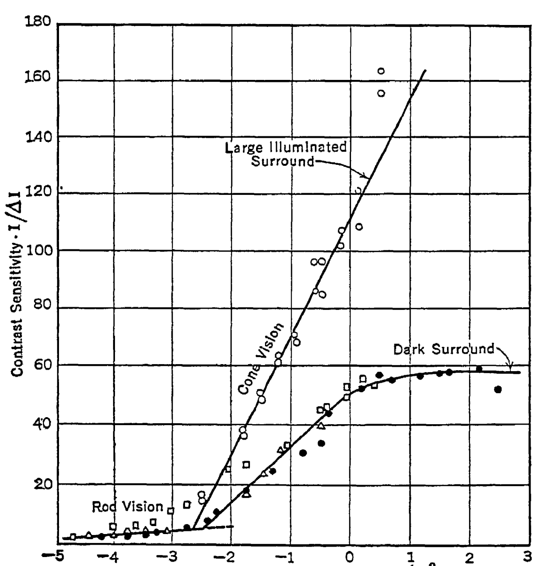


FIG. 8. The Contrast Sensitivity of the Eye, Showing Data of Koenig, Blanchard, Holliday, and Stiles and Crawford (after Moon, *Scientific Bases of Illuminating Engineering*)

part of the curve the fraction $\Delta I/I$ is nearly constant, and only in this region can the differential threshold of the eye be said to illustrate Weber's law. For very high luminosities a slightly decreased sensitivity may be found, but this can be attributed to uncontrolled illumination of the more peripheral surroundings of the test areas. That part of the curve below 0.013 millilambert results from rod vision only, and a somewhat abrupt change near that point is suggested by the data. It is sometimes more instructive to plot the differential threshold data as the contrast sensitivity ($I/\Delta I$), as shown in Fig. 8.

If smaller test and standard areas are used, the sensitivity is less, that is, the curve lies higher on the $\Delta I/I$ axis of the differential threshold graph. Also it is found that, if the luminosity of the surroundings outside of the test and standard areas is kept constant, and then that of the standard area is varied, the differential threshold is increased (sensitivity decreased), and the entire curve lies above the curve shown in Fig. 7. The luminosity of the surroundings of the test field is an important factor in the contrast sensitivity of the eye.

ADAPTATION. When one goes from a brightly illuminated room to one that is only dimly illuminated, several minutes must elapse before details in the room can be discerned. Likewise in going from the dark room into sunlight there are a few moments of blinding glare. In either case, the eyes become adjusted to the prevailing brightness of the visual field in a few minutes. The retinal process by which this occurs, as well as the final stationary state, is called *adaptation*. By the process of adaptation the sensitivity of the eye to contrast differences is automatically changed to meet changing luminosities of the surroundings. In light adaptation (photopic vision) the visual sensitivity of the eye decreases; in dark adaptation (scotopic vision) it increases. In this adaptation the eye adapts itself to changes of illumination of several thousand times.

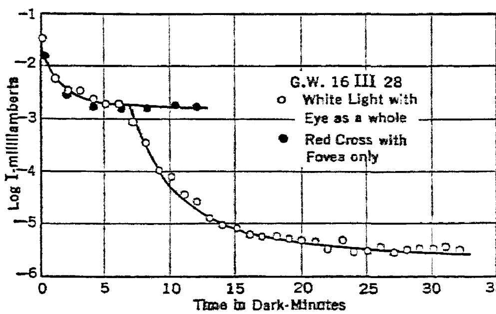


FIG. 9. The Dark-adaptation Curve of the Eye as a Whole and of the Fovea Showing the Behavior of the Cone and Rod Systems (Hecht, *J. Gen. Physiol.*)

Light and dark adaptations are opposing processes and are different for the cone and rod systems of the retina. To a small extent the pupil aids in the adaptation processes. A typical curve showing the progress of dark adaptation of the eye as a whole is shown in Fig. 9. The ordinates indicate the luminosity of the test object at the threshold of visibility at any given moment in the progress of the dark adaptation. The upper curve results from the activity of the retinal cones; the remaining part of the curve results from the rods. The data show that foveal (cone) dark adaptation is nearly complete in 2 to 10 min. For the rods the eye is essentially dark adapted in 40 min, though the sensitivity can still be shown to increase slightly for longer periods of time in total darkness. Anomalies in dark adaptation may be due to vitamin A deficiency.

The progress of dark adaptation varies somewhat with the size of the test object and especially with the intensity of the illumination in the previous light adaptation. There is evidence also that the rate of dark adaptation can be increased and the dark adaptation maintained when the observer wears dark red glasses, adequately shielded, in ordinary illuminations (Rowland and Sloan).

The rate of light adaptation is very much more rapid than dark adaptation. Though dependent upon the intensity of the adapting light, the sensitivity drops to a fraction of its initial value within the first few seconds, and the light adaptation is nearly complete in 1 min.

10. TEMPORAL ASPECTS OF PERCEPTION

PERSISTENCY OF VISION. From the instant the retina is stimulated to the moment the first sensation occurs, a brief interval of time has elapsed. This interval, called the *latent period*, which, in a sense, is a visual reaction time, depends upon the nature and strength of the light stimulus and upon the adaptation of the retina. Under ordinary light conditions, the latent period is roughly between 0.06 and 0.2 sec. As the intensity of the stimulus is increased, however, the latent period decreases to a minimum (0.065 to 0.130 sec), beyond which further increases in intensity will not shorten the latent period.

Above the minimum, this interval varies approximately inversely as the logarithm of the intensity. In dark adaptation, the latent period, for threshold conditions, may be 0.5 to 1 sec, depending upon conditions. The latent period is generally shorter in the peripheral parts of the retina than at the fovea, and shorter for blue light than for red.

With a flash stimulus, the sensation persists for some time after the stimulus has ended. The magnitude of the light sensation and its duration vary with the adaptation level of the eye, that is, the luminosity of the background upon which the stimulus is seen and the strength of the stimulus. Typical curves are shown in Fig. 10. The duration of the impressions from stimuli of equal brightness is longer for the fovea than for the peripheral parts of the retina and is generally increased as the eye is more dark adapted. An increased intensity of the stimulus results in a decreased duration time. For colored light the persistence of the sensation varies inversely to the apparent luminosities of the visibility curves and hence is shortest for yellow, longest for blue, and intermediate for red.

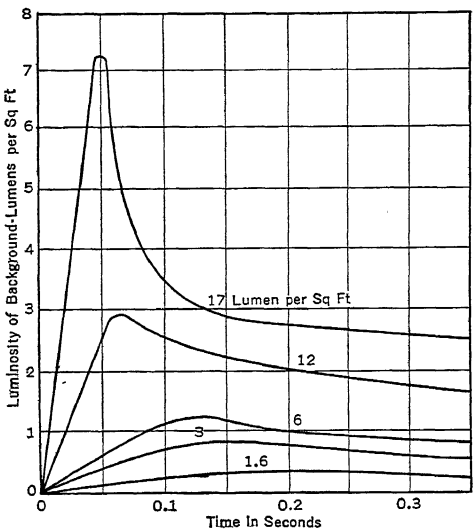


FIG. 10. Illustration of the Apparent Brightness and the Duration of the Light Sensation Following a Flash Stimulus (Data of Broca and Sulzer, from Luckiesh and Moss, *Science of Seeing*)

After the first sensation has faded away, several after-images may again appear at intervals depending upon the luminosity of the surroundings and the intensity of the stimulus. If the after-image corresponds to the original impression as regards contrast and color it is called positive; if the light-to-dark relationships are reversed and the colors are complementary to the original impression it is called negative. Considerable practice is sometimes necessary to see these images, for conditions of fatigue, adaptation, and inhibition may greatly affect their appearance. The typical sequence of after-images is illustrated in Fig. 11, where the ordinate represents the relative brightness of the after-images, positive above and negative below the abscissa.

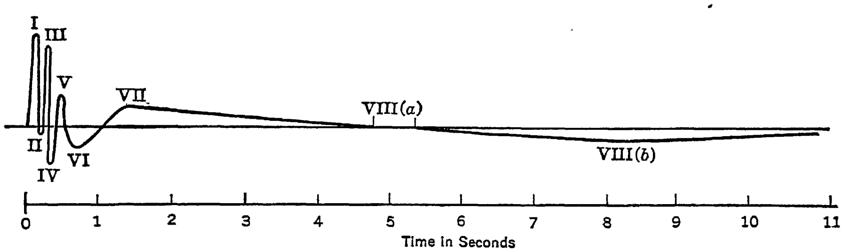


FIG. 11. Diagram of the Sequence of After-effects Following a Flash Stimulus in the Light-adapted Eye (after Tschermak-Bethe, *Handb. normal u. path. Physiol.*, 12/1, II)

PERIODIC STIMULI AND FLICKER. Periodic light stimuli may be perceived as discrete flashes, but if made sufficiently rapid, so that the persistent sensation from one stimulation curve overlaps the rise of the primary sensation from the succeeding stimulation, the sensation will be the same as that for a continuous illumination.

The brilliance of fused periodic stimuli which have different intensities is the same as the average intensity (Talbot's law), that is, the integral of the photometric luminosity of the periodic stimuli divided by time. This law holds accurately except for extremes of high and low intensities. That frequency at which periodic stimuli are first perceived as a steady illumination (fused) is known as the *critical flicker frequency*, which may be abbreviated to c.f.f. The critical flicker frequency varies with the illumination, the part of retina being stimulated, the area of the flickering field, the ratio of the light-to-dark intervals of the flashes, the retinal adaptation, the wavelength of the light, and the presence of other steady light stimuli also falling upon the retina. So reliable is the c.f.f. under controlled conditions that it is sometimes used to measure luminosities and adaptation levels.

The manner in which the c.f.f. for small test fields illuminated with white light varies with illumination and for different parts of the retina is shown in Fig. 12. In the rod-free area at the fovea, the c.f.f. varies proportionately with the logarithm of the illumination (Ferry-Porter law), except at the extremes of the illumination. The maximum critical frequency is about 53 cycles per second. The c.f.f. decreases for flicker stimuli on the peripheral parts of the retina. If the area of the flickering field is large, however, the

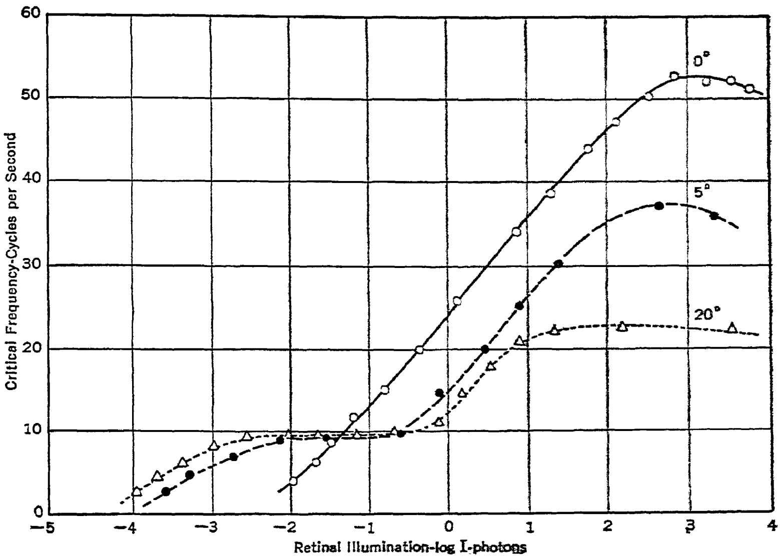


FIG. 12. The Relationship between the Critical Flicker Frequency and the Illumination, for Different Parts of the Retina (Hecht, *J. Gen. Physiol.*)

critical frequency may be found higher in the periphery, which indicates a spatial summation of the retina in the periphery. The curves generally appear to be in two parts, a division which is attributed to the rod and cone behavior. The c.f.f. for various colored stimuli is substantially the same as for white light if their apparent luminosity is the same as that of the white light and the intensity is above the cone threshold.

In the curves shown in Fig. 12, the duration of the light phase was equal to that of the dark. Where the dark phase is longer, generally higher critical frequencies are found, and when shorter, lower critical frequencies are necessary, especially in the ordinary ranges of illuminations. Only at high intensities of the light stimuli will this generalization be invalid. Rapid eye movements or rapidly moving periodic stimuli enhance the appearance of flicker, and under these conditions higher frequencies are necessary. A steady illumination of one part of the retina changes the c.f.f. of a periodic stimulus in another part if the separation of the two stimuli is not too great. In general the c.f.f. increases as the brightness of the steady stimulus is increased, until a maximum is reached, beyond which it then decreases.

Bartley found that for periodic stimuli below the c.f.f., and at about 8–10 flashes per second, there is a marked enhancement of the apparent brightness of the stimuli.

11. COLOR

A clear differentiation must be made between the *color stimulus* and the *color sensation.* The color sensation and all the associated phenomena are purely psychological (experiential). The stimuli measured in terms of energy and wavelength are physical. The color problem is fundamentally the finding of the relationships that exist between the color sensation and the physical stimuli. Moreover, relationships found for color mixtures from spectral light will not necessarily be true when colored pigments are mixed. This fact has led to much confusion and controversy. The spectral luminosities reflected from pigments, however, are subject to the same laws as those from spectral light.

Colors are ordinarily seen as properties of objects and hence are usually associated with the objects themselves. There are, therefore, many psychological constancy phenomena in which objects tend to retain their color and brightness when seen in varying conditions of light and shade. Photographic scenes in color projected in an otherwise darkened room tend to retain normal color relationships in spite of actual wide deviations. Under other conditions where these reproductions are viewed against backgrounds of stable color and contrast relationships, these deviations are more readily apparent. To study correlations between the quality of color experience and the physical composition of light, the test fields must be dissociated from known objects having spatial values. These psychological facts are important in dealing with color problems.

A given color sensation is said to have three dimensions: *hue, saturation,* and *brilliance.* Hue is associated with the dominant wavelength of the light stimulus. Saturation indicates the amount or degree of the hue present (deep-red as against pale-red) and represents the ratio of the luminosity of the pure spectral light to that of the white light present. Brilliance (apparent brightness, apparent luminosity) indicates the total intensity (energy) of the colored light stimulus. Frequently, hue and saturation are considered together under the term *chromaticity.*

SATURATION OF THE SPECTRAL COLORS. All spectral colors do not appear equally saturated, that is, some have a greater sensation of white than others. The ratio of the luminosity of the least perceptible spectral color to the luminosity of a background field defines the least-perceptible colorimetric purity for that color. Two halves of a test field are equally illuminated with white light. To one half a spectral color is added, and, in order that the brightness of the two fields shall be the same, the luminosity of the white light of that field is decreased as the spectral color is added. The intensity of this spectral color is then increased until the two halves of the field just appear different in color. From the luminosity of the added spectral colors the colorimetric purity of the spectrum can be found. Figure 13 illustrates data obtained in this manner. It is clear that a great deal

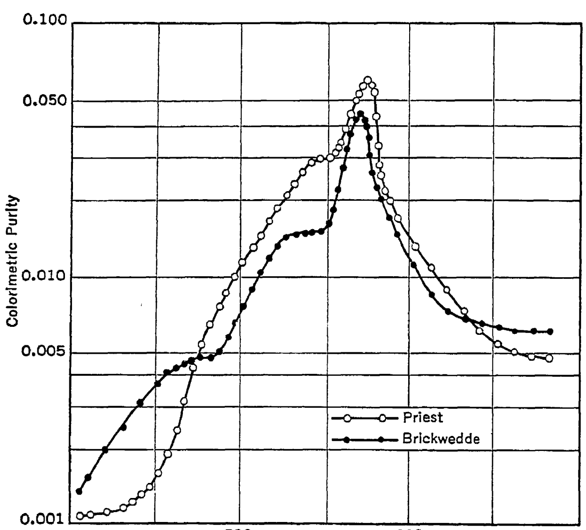


FIG. 13. The Least-perceptible Colorimetric Purity of the Visible Spectrum (Data of Priest and Brickwedde, after Hecht, *J. Gen. Physiol.*)

more of yellow than red or blue is required to produce a least-perceptible color, and yellow is therefore considered a less saturated color than red or blue.

HUE DISCRIMINATION THRESHOLD. The change in wavelength corresponding to a just-noticeable difference in color varies over the spectrum and also somewhat with the individual. In general, however, the data are similar, and the resultant curves show three maxima and three minima. In Fig. 14 are illustrated representative data for an individual with normal color vision. The discrimination is poorest at the ends of the spectrum, especially in the red.

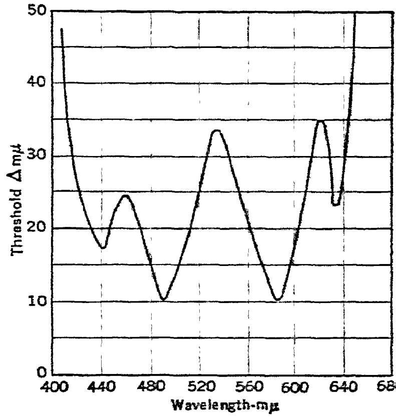


FIG. 14. The Least-perceptible Difference in Color for the Visible Spectrum (Data of Jones, *J. Optical Soc. Am.*)

COLOR SPECIFICATION. The color stimulus (as distinct from the color sensation) of any color field can be specified quantitatively by the luminosity of the radiation given off for each part of the spectrum. This can usually be measured by a spectroradiometer or a spectrophotometer. Under any condition, the objective color from a given surface can then be specified by a spectrophotometric curve obtained from those measurements.

The eye itself cannot analyze the radiation from a color stimulus as can the spectrophotometer, for it responds in a complex manner dependent not only upon the visual processes and their reaction to light stimuli but also upon certain psychological factors. There are, therefore, many different objective color stimuli, as specified by spectrophotometric curves, which will result in the same color sensation. It is possible, however, to transform the data from the spectrophotometric measurements by means of standard data from color-matching experiments, so that equal color sensations can be specified by three quantities derived from the objective stimulus.

The I.C.I. (International Commission on Illumination) system is that most used for accurate color specification. It is based upon the theory that the chromaticity (hue and saturation) of any color stimulus can be specified by three quantities, which represent the proportions of three selected primary light distributions that are necessary to match the color sensation evoked by a given stimulus. The amounts (luminosities) of each of three selected primary color sources that must be added together in order to match a given spectral color are determined by these color-matching experiments. These amounts are called the *tristimulus* values of the three primaries for that color. Three standard spectral distributions of equal energy, whose tristimulus values are $\bar{x} = f_x(\lambda)$, $\bar{y} = f_y(\lambda)$, and $\bar{z} = f_z(\lambda)$ (λ being wavelength), are especially selected, with dominant wavelengths in the red, green, and blue respectively. Cf. Fig. 15. These primary distributions are obtained by transformations from the experimental data with real primaries; they are specially selected to avoid negative tristimulus values and to meet other requirements for a convenient system. The $\bar{y}$ spectral distribution function has been made to correspond to the photopic visibility curve. Any homogeneous spectral radiation can be specified by the tristimulus values read directly from the curves or special tables (Hardy). For a spectral radiation objectively specified by a spectrophotometric

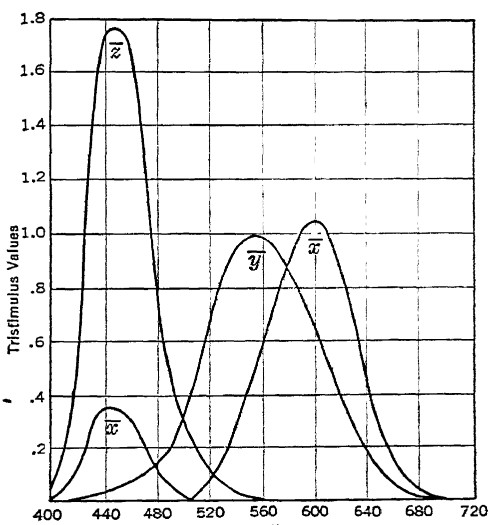


FIG. 15. Tristimulus Values for the Various Spectral Colors. The values $\bar{x}$, $\bar{y}$, $\bar{z}$ are the amounts of the three I.C.I. primaries required to match in color a unit amount of energy having the indicated wavelength. (Hardy, *Handbook of Colorimetry.*)

curve, the tristimulus values are found by summing up the products of the energy E_λ for all wavelengths by the corresponding tristimulus values from the standard primaries from the tables, viz.,

$$x' = \int_0^\infty \bar{x}E_\lambda \, d\lambda \qquad y' = \int_0^\infty \bar{y}E_\lambda \, d\lambda \qquad z' = \int_0^\infty \bar{z}E_\lambda \, d\lambda$$

Through the choice of primary standard distributions the second integral gives directly the relative luminosity (brightness) of the color stimulus on a black (zero) to pure white (100) scale. In practice, the tristimulus values for any spectrophotometric curve are computed by averaging the products of $\bar{x}E_\lambda$ (and $\bar{y}E_\lambda$ and $\bar{z}E_\lambda$) for equally spaced wavelengths.

The values of x', y', and z' do not necessarily measure the color sensation, but they do state the conditions under which different spectral stimuli will result in the same color sensation. It must be borne in mind that this sensation may vary slightly between individuals and even with other factors, such as size of field and conditions of the surrounding fields.

THE CHROMATICITY DIAGRAM. A convenient graphical representation of the chromaticity of any spectral stimulus is made by plotting the trichromatic coefficients which are defined by the ratios

$$x = \frac{x'}{x' + y' + z'} \quad \text{and} \quad y = \frac{y'}{x' + y' + z'}$$

and the third would be related to the other two by $x + y + z = 1$. The result is the standard chromaticity diagram in which colors of equal luminosity are represented in terms of dominant color and saturation. See Fig. 16. Color matches are known to be stable through wide limits of luminosity; hence, this representation is valid.

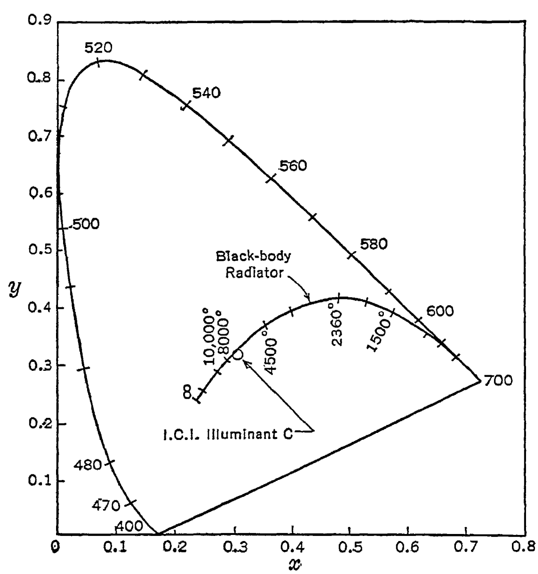


FIG. 16. The Chromaticity Diagram, Showing the Geometrical Arrangement of the Spectral Colors and the Locus of White Lights (after Hardy, *Handbook of Colorimetry*)

The visible spectrum arranges itself in a horseshoelike curve beginning with red (700 mμ) and passing counterclockwise through orange, yellow, green, blue to a deep purplish blue (400 mμ). The straight line joining the ends of the spectrum contains the purples. Clearly, this line is not part of the spectrum. White light is represented by various points on a curve near the center of the diagram, its exact position depending upon the temperature of the source. The luminant recommended by the International Commission on Illumination is shown on the diagram at C. Any given color stimulus will also be represented by a point within this curve. Figure 17 shows the approximate color names for the different portions of the diagram. The dominant wavelength of any stimulus will be given by the place where a straight line drawn through C and the point intersects the nearest part of the spectral curve. The complementary wavelength will be the intersection point on the spectral curve diametrically opposite. The saturation or excitation purity of a given color can be determined from the ratio of the distances of the point from C, and from the dominant wavelength on the spectrum curve. The resultant color of a mixture of any two colors represented by two points on the I.C.I. diagram will be represented somewhere on the straight line drawn between the two points. The advantage of this representation is its direct quantitative application to spectrophotometric data. Its disadvantage, however, lies in its graphical distortion of color differences.

OTHER SYSTEMS. The Ostwald and Munsell systems consist of orderly arrangements of colors based upon their visual relationships, and not dependent upon the mixture of pigments or upon the physical measurements of the light, and independent of the illumination. Both these systems are illustrated in carefully prepared charts with orderly arrays of colored patches, which vary in hue, saturation, and value according to the concept of the author.

COLOR TOLERANCE. Nearly as important to color specification is the necessity of knowing color tolerances. The most thorough study, to date, is that of MacAdam in which the color tolerances in changes in hue and in saturation were determined.

COLOR ADAPTATION. If the eyes are subjected to a large field of a spectral color for a period of time, the perception of hue of other colors becomes modified and distorted. This may be in the nature of an adaptation. Two electric lamps, one white and one colored, say red, are set in front of a large white screen several feet apart. Then an object is placed nearer the screen in between the lamps, so that two shadows, one from the white and one from the red lamp, are cast upon the screen. Upon continued observation it will be found that the shadow cast from the red lamp will appear green, complementary to the red color, while the shadow from the white light will appear red. In the general perception of colored objects in fields of familiar detail, the color relationships tend to remain the same, in spite of rather wide variations of illumination intensities and even color. If the visual fields are isolated and confined, however, to small areas, apart from the larger field, this color-constancy phenomenon tends to disappear.

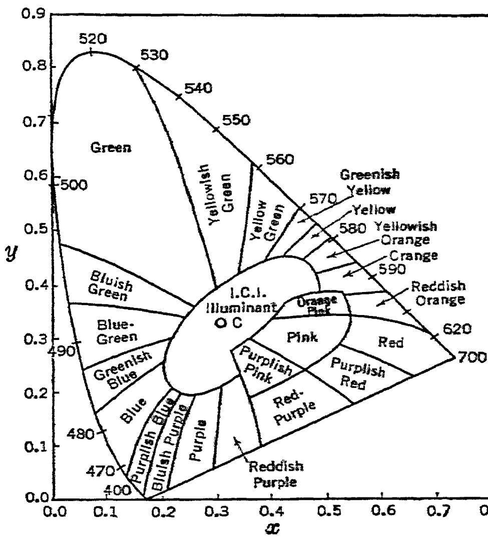


FIG. 17. Approximate Color Names for Various Parts of the Chromaticity Diagram (Kelley, *J. Optical Soc. Am.*)

12. THE SPACE SENSE

It is necessary to distinguish clearly between objective space, filled with real objects, and *visual space*, filled with *visual objects*. It is somewhat meaningless to ask whether the position of visual objects is identical to that of the real objects. It is only necessary that the relationships between visual space and objective space be sufficiently stable so that the organism can effectively operate within real space. The visual perception of space is essentially egocentric in that objects are localized with respect to the body. The final perception of space arises through a complex integration of (1) the visual clues inherent in the pattern of the dioptric images that fall on the retinas, (2) the simultaneous impressions from the other senses, (3) experiential factors that have been associated with visual, auditory, and tactile stimuli, and (4) the immediate attention of the individual. It is convenient to regard space perception in terms of (1) the discrimination of direction (egocentric and bidimensional), and (2) the discrimination of depth, that is, the third dimension.

DIRECTION LOCALIZATION. By virtue of the discrete make-up of the retina (the retinal mosaic), the various parts of the dioptric images on the retina can be differentiated through some type of "local signs" associated with the retinal elements, which, in this case, is a subjective visual direction. Resolving power and visual acuity are measures of the keenness with which this differentiation can be made. The fovea is the primary point of reference, and the subjective direction of objects is described in terms of "breadth" (right or left) and "height" (above or below) the point of fixation which is imaged on the fovea. The subjective visual directions then are relative to the fovea. By a reflex process associated with the eye, head, and body movements, the change in the subjective direction of the whole visual field that would occur with eye movements is counteracted, so that objects tend to appear in the same "absolute" direction in spite of the fact that the images move across the retina.

The precision of the relative subjective direction varies greatly with the character of the visual field, and in many cases estimations actually are inaccurate. Probably the highest precision is in the estimation of when lines are parallel or when they are straight, and in the comparison of angles whose corresponding sides are parallel. In the estimation of differences in length, a greater precision is found in the horizontal than in the vertical. Vertical distances, moreover, look longer than equal horizontal distances. In the estimation of lengths, accuracy varies with the lengths to be compared and their relative positions. When attempting to determine the center of line segments without eye movements, the eye tends to overestimate the portion on its own side (Kundt partition phenomenon), though frequently the reverse is found. Subdivided and filled spaces look longer than unfilled spaces. Acute angles are generally overestimated and obtuse angles are underestimated. Magnitudes are diminished in the presence of larger magnitudes and magnified in the presence of small ones. Many of the well-known illusions are examples of these errors in the perception of direction.

VISUAL ACUITY AND RESOLVING POWER. The *visual acuity* of the eye is the degree to which it is able to discriminate fine detail in the visual field, and this varies considerably with the type of detail, the contrast, illumination, surrounding brightness, adaptation level, etc. In a strict sense, *resolving power* is more definitive than visual acuity; it is measured as the reciprocal of the smallest visual angle (usually in minutes of arc) by which two objects can be seen separately. The theoretical limit of visual resolution would be determined by the size of the diffraction circles on the retina and the dimensions of the retinal elements. However, for pupils of normal and larger sizes the spherical and chromatic aberrations, together with small irregular astigmatism that is usually present, modify the nature of the light intensity distribution within the image, so that, with the contrasts ordinarily encountered, the resolution is better than that based upon diffraction alone. For small pupils a marked decrease in acuity is found, probably resulting from the increased size of the diffraction circles, but for pupils larger than 2 to 3 mm the acuity remains nearly constant (Cobb). Faint stars whose angular separation is 1 to $1\,{}^1/_2$ minutes of arc can usually be resolved. With small point light sources of low contrast with respect to the background, the minimum angle of resolution may be 100 seconds with a mean error of 5 seconds of arc. With repeating patterns such as lattices, grids, or checkerboard patterns the minimum angle of resolution varies between 50 and 75 seconds of arc, under optimum conditions.

The eye is, of course, capable of much finer discrimination of detail than that which would be obtained on the basis of resolving power alone. Experiments which indicate this fine appreciation of detail involve least-perceptible differences in contours. These differences have been expressed as the mean error of settings, or in the 50/50 point, in correct judgments. Typical results are shown in the table below.

Test	*Mean Error, seconds*	*Authority*
Widening of lower half of slit	10–12	Wulfing (1892)
Coincidence of vertical lines (the vernier)	8–12	Brian and Baker (1912)
Coincidence of vertical lines (the vernier)	13	Best (1900)
Coincidence of vertical lines (the vernier)	3	Langland (1929)
Coincidence of vertical lines (the vernier)	0.6	French (1920)
Coincidence of vertical lines (the vernier)	2	Wright (1942)
Error in contact of white disks on dark background	15	Dale (1920)
Alignment of edges of two rectangles	10	Hering (1899)
Error of setting range finders	2–6	von Hofe (1920)
Stereoscopic displacement of images	5–7	Florian (1930)

In general, where such fine discrimination can be seen, the retinal images of the detail (such as extended lines) involve the activity of a larger number of retinal elements. Except for monochromatic yellow light, visual acuity is decidedly poorer under colored illuminants than under white light. This is especially true of blue and only slightly less so under red illuminations. Part of this latter decrease in acuity is undoubtedly related to the chromatic aberration of the eye.

Visual acuity is usually measured by a test chart of letters of graded sizes, in spite of certain obvious inherent faults. Visual acuity is considered normal when letters can be identified the separation of adjacent parts of which subtend an angle at the eye of 1 minute of arc. A line of letters, each of which subtends a 5-minute visual angle, is printed on a chart in sequence for visual distances of 10, 15, 20, 25, 30, 40, 60, 100, and 200 ft. These letters are usually arranged with the smallest at the bottom of the chart, and with a single large E at the top. The acuity tests are usually made at a visual distance of 20 ft, and that line of letters which the subject can just read is then recorded relative to 20 ft. Thus 20/20 represents normal vision; 20/15 indicates that print which could ordinarily be read at 15 ft can be read at 20 ft and therefore the acuity is better than normal; and 20/100 indicates that that type which should be legible at 100 ft can be read only at 20 ft and therefore the acuity is much less than normal. Other test objects frequently used are the Snellen hook, the Landolt broken ring, and checker-board patterns.

Because of the increased size of the photosensitive elements toward the peripheral parts of the retina and the greater number of them associated with single conductor nerves, together with the increased magnitude of the optical aberrations toward the periphery of the eye, it is to be expected that the resolving power of the eye would fall off rapidly toward the periphery. Figure 18 shows this decrease in resolving power with the increase

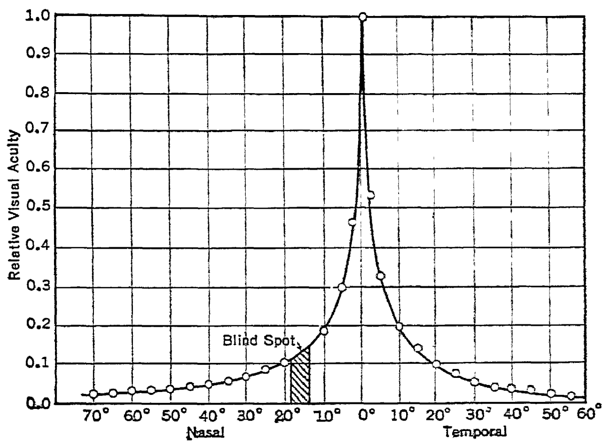


Fig. 18. The Decrease in Visual Acuity toward the Peripheral Parts of the Retina (Wertheim)

of the peripheral angle (data of Wertheim). A very rapid decrease in acuity occurs to about 5 degrees, and from there the decrease is much slower. This loss of acuity is not the same in all the meridians of the eye but is greater in the vertical meridian than in the horizontal. Figure 19 shows the isopters of the retina, that is, the curves of equal visual acuity (data of Wertheim).

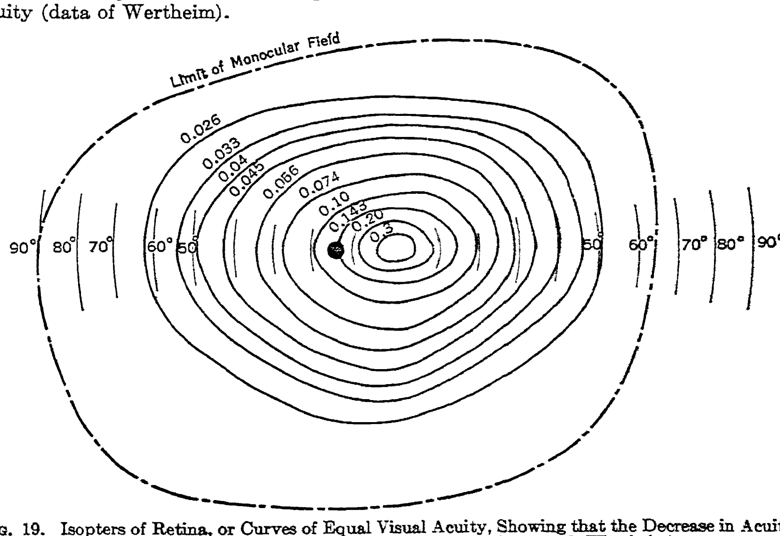


Fig. 19. Isopters of Retina, or Curves of Equal Visual Acuity, Showing that the Decrease in Acuity is Greater in the Vertical Meridian than in the Horizontal (Wertheim)

There has been evidence that visual acuity is greater for distant vision than for near vision. This phenomenon is believed to be dependent upon the type of test and the conditions under which the test is made, especially for peripheral vision, because opposite results have been reported.

IRRADIATION. Owing to the aberrations of the optical system of the eye, the images of discrete points are not defined with sharp boundaries. Rather, the intensity of the light in the image falls off in a bell-shaped curve. The perceived contour depends upon the differential sensitivity of the eye, but the image always corresponds to an object larger than the original. The position of the perceived contour will tend to be toward the less intense end of the light distribution curve. Accordingly, bright objects seen against a dark background appear larger than dark objects of equal size against a bright background. This phenomenon, called *irradiation*, depends upon the relative luminosity of the adjacent surfaces and varies with individuals because of differences in ocular aberrations. Objects of large angular size are increased proportionately only to a small extent, and their size can be said to remain constant. For small objects, however, the irradiation increases with decrease in angular size. On the other hand, the separation of small black lines against a bright background may actually appear larger than it is. This so-called negative "irradiation" is explained, however, by the influence of the mechanism of *simultaneous contrast*, whereby the apparent brightness (or color) of an area is influenced (enhanced) by adjacent areas of different brightness (or color).

VISUAL ACUITY AND ILLUMINATION. The visual acuity of the fovea of the eye increases markedly with an increase in the luminosity of the field, but that of the peripheral retina changes scarcely at all. In general, then, studies on visual acuity and illumination pertain to foveal vision. The data of König illustrated in Fig. 20 (as recomputed by Hecht) are among the most complete on this phenomenon. They were obtained with a black test object upon a white background and are therefore of maximum contrast. For illuminations from 0.013 to 10 lumens per sq ft the visual acuity is approximately proportional to the logarithm of the luminosity. The visual acuity obtained at lower illuminations decreases rapidly in the neighborhood of 0.001 lumen per sq ft, where scotopic vision begins. From here, for lower illuminations, only rod vision is involved and the change in visual acuity is small. Although König's data show a maximum for illuminations greater than 100 lumens per sq ft, and thereafter a constant visual acuity, more recent data have indicated that this maximum is due to the fact that the brightness of the field surrounding the background of the test object was low compared to that of the test object itself. In experiments where the brightness of the surrounding field is the same as that of the background for the test object, a maximum is not found and the visual acuity continues to increase at only a slightly lower rate with a continued increase of illumination. The visual acuity is lower for colored light than for white of the same brightness, being poorer in the order green, red, blue. In general, visual acuity is also reduced under conditions of glare.

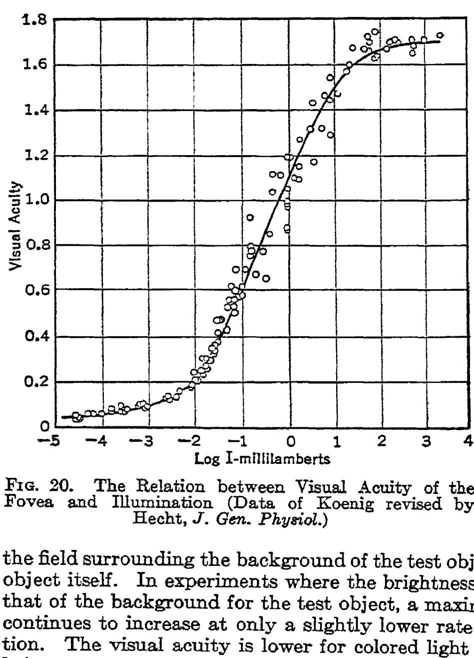


FIG. 20. The Relation between Visual Acuity of the Fovea and Illumination (Data of Koenig revised by Hecht, *J. Gen. Physiol.*)

When detail is exposed for short intervals of time (less than about 1 sec), and the luminosity of the background remains constant, the visual acuity varies roughly with the logarithm of the time of exposure. Below exposure times of about 0.01 to 0.03 sec, depending upon the illumination, the product of time and illumination is constant for the same visual acuity.

In König's data the contrast, defined as the ratio of the difference in luminosity of the background and the test object to the luminosity of the background, was very high. It has been shown that with lower contrasts the visual acuity is also lower. Figure 21 illustrates data from Cobb and Moss on the relationship between contrast and visual acuity. No data are available for bright detail against a darker background.

MINIMUM VISIBLE. Similar to the problem of visual resolution is that of the minimum visible. What is the angular size of the smallest detail that can be perceived? As in resolving power, this depends upon an intensity discrimination within the image on the retina and hence varies with the nature of the dioptric image, the contrast sensitivity of the eye, the size and form of the detail, exposure time, etc. Lines are more readily seen than dots, and repeating patterns of dots, lines, etc., are perceived even more readily. Small voluntary and involuntary eye movements also aid in the discrimination of fine detail. Under general illumination a line against a bright background can be seen when its width subtends a visual angle of 3 to 4 seconds of arc. Hecht and Mintz found that with a similar test object the angular width of a wire that could just be seen varied from 10 minutes to 0.5 second over the complete range of illuminations. The minimum value 0.5 second represents a discrimination of about 1 to 2 per cent difference in light intensity in the image on the retina. When a narrow illuminated slit that is exposed for short durations is just visible, it is found that (visual angle) × (duration of exposure) × (adapting luminosity of background) is approximately a constant for flashes of 0.004 to 0.189 sec of duration (Niven and Brown).

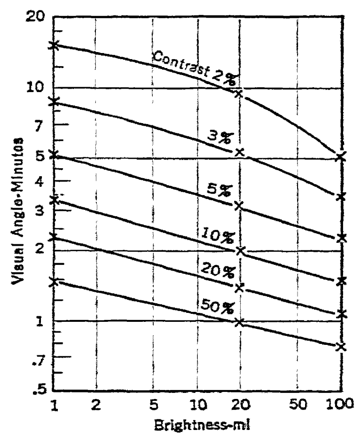


FIG. 21. Relationship between Minimum Angle of Resolution, Luminosity of Background, and the Contrast of the Test Object Relative to its Background (Cobb and Moss, *J. Franklin Institute*)

As indicated above it is now believed that the rate at which contrast sensitivity and visual acuity in the absence of glare increase with luminosity of background is not appreciably diminished with the very high luminosities (cf. Figs. 8 and 20). In other words, there is no optimum luminosity. The question arises then as to what the general illumination should be for adequate vision. Although this question cannot be answered specifically

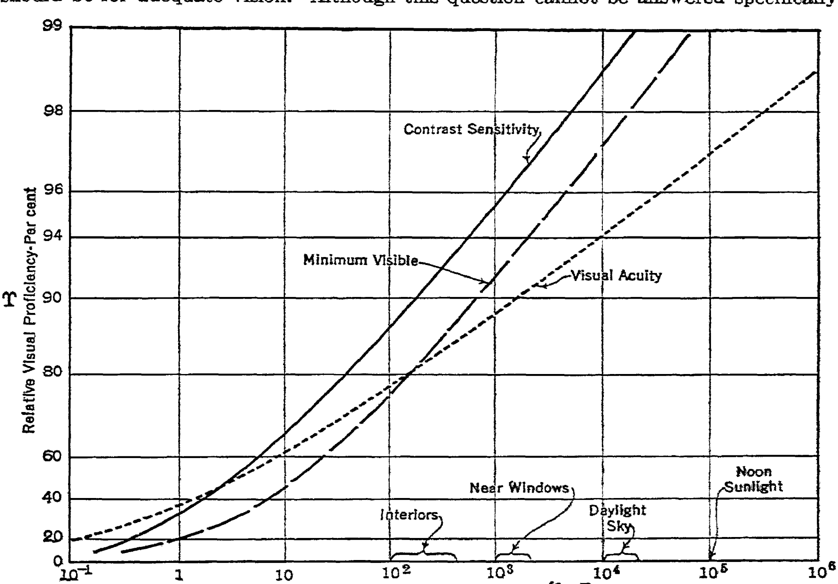


FIG. 22. Relative Visual Proficiency for Contrast Sensitivity, the Minimum Visible, and Visual Acuity, as a Function of Brightness of Background (after Moon and Spencer, *J. Optical Soc. Am.*)

it is of value to know the relative degree of visual efficiency for any given general illumination. A scheme for indicating this has been suggested by Moon and Spencer. In contrast sensitivity, the quantity Υ_c is defined as the ratio of the contrast sensitivity

$(I/\Delta I)$ for the given illumination (I) to the maximum contrast sensitivity that would be obtained with uniform fields and very high daylight luminosities. It represents the relative proficiency of contrast discrimination at any given illumination. Similar ratios Υ_v and Υ_a are defined for the minimum visible and for visual acuity. In Fig. 22 the theoretical values of each of these three visual functions are shown as dependent upon the luminosity of the background. For outdoor lighting the relative visual proficiency would range between 70 and 95 per cent, and for interiors with artificial lighting it would probably be less than 80 per cent.

PERCEPTION OF MOTION. The motion of objects in space is perceived as motion relative to the observer or as relative motion between the objects themselves. Under special circumstances the motion of objects seen with respect to each other and to the observer may be equivocal, for example, the apparent motion of the moon observed through moving clouds. In a field of large and small objects the larger objects are said to exhibit more stability and are less liable to apparent motion, while small objects exhibit apparent movement more readily and are said to be more mobile. Moreover, objects imaged on the peripheral parts of the retina appear less mobile than those falling on the fovea.

Perception of motion rests upon a temporal reaction to stimuli falling successively on neighboring points of the retina, but its appreciation appears only partially related to resolving power. It is thought to be a visual sensation resulting from experience. For objects whose visual angle of movement is small, motion is inferred from the apparent positions after lapses of time, for example, a moving train seen at a great distance. The true perception of motion, however, is a specific sensation, with both lower and upper thresholds of discrimination. The impression of motion can also occur (within definitely prescribed limits) from successive stimuli arising from separated stationary sources. This is called apparent movement by psychologists.

The threshold of the perception of true motion relates to the least angular displacement of an object that can be recognized in unit time. With comparison objects in the field of view the lower threshold is 1–2 minutes of arc per second. Without comparison objects this value must increase 10 to 20 times. The least angular movement that can be detected between two fixed points is sometimes called the movement acuity; under ordinary circumstances, this is found to be 10 to 20 seconds of arc, when stationary comparison objects are in the field of view. Without comparison objects this increases to more than 1 minute of arc. When the range of the movement (the distance between the beginning and end points) is fixed, the shortest lapse of time in which motion is seen has also been measured. Under ordinary circumstances, for example, the threshold for a 10-degree range varies between 0.027 and 0.079 sec. The upper limit for the discernment of motion, when the sensation becomes a meaningless blur, is 1.4 to 3.5 degrees per $^1/_{100}$ sec. The movement acuity decreases with increase in illumination.

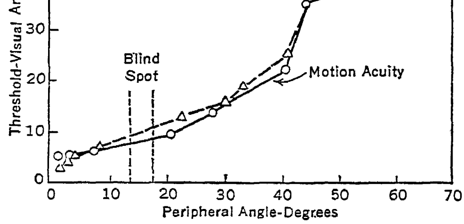


FIG. 23. The Thresholds for Form Acuity and Motion Acuity for Same Test Objects. The motion acuity is given in terms of the minimum distance required for perception of motion. (Data of Klein, *Arch. Psychol.*)

The threshold for the discrimination of movement is least at the fovea and increases rapidly out toward the periphery of the retina, though it has been thought that the increase is much less than the decrease of visual acuity toward the periphery. Figure 23 illustrates the data of Klein, which show, however, that the acuity and motion thresholds are essentially the same except at the fovea. It appears that the *discrimination* of motion is greatest at the fovea, but the *appreciation* of motion is most marked in the peripheral retina, to a degree beyond the actual resolving power. The discrimination of movement in the periphery is somewhat poorer above and below the fovea than to the right or left. There is a tendency to overestimate the extent of motion in the periphery, and, furthermore, the subjective direction of that motion is somewhat vague and uncertain. However, the sense of movement is highly developed.

THE PHI-PHENOMENON. The impression or illusion of movement can occur in certain situations without there being actual physical movement. The most important of the apparent movements is known in its purest form as the *phi*-phenomena, whereby apparent movement can occur, under certain conditions, when separated stationary points are illuminated successively. Usually the illusion of movement can be distinguished from real motion, however, and, as such, may be variously interpreted by different observers. Only in the neighborhood of an angular succession of 20 degrees per second are the two motions difficult to identify (de Silva). For the impression of motion there is a time relationship between the duration of the first stimuli, the separation of stimuli, the interval between successive stimuli, and the brightness of background, the impression depending also to a great extent upon the attitude of the observer. Under ordinary conditions, with a dark background, it has been found that, for two point stimuli 1 cm apart, the duration of each of which is 0.05 sec: (1) if the two stimuli are nearly instantaneous (0.03 sec) they will appear simultaneous; (2) if the time interval is greater than 0.2 sec the points will appear discrete and stationary; (3) for time intervals intermediate between these two there will be an apparent movement of the first point to the second, with an optimum at 0.06 sec. Of two light points of different intensity the weaker tends to move toward the brighter. Other examples of apparent movement are in the cinema, in neon signs, and in stroboscopic instruments.

THE PERCEPTION OF DEPTH. In the perception of depth one must distinguish between the perception of depth differences (how much one object appears in front of or behind another) and the absolute localization by which the actual distance of objects from the individual is estimated. Binocular vision, through the phenomenon of stereopsis, provides the most accurate means of relative depth discrimination.

The absolute localization of objects in space results from a complex process that involves both monocular and binocular perception. It has been usual to state that the perception of depth by monocular vision is a conception of depth (distance) attained through experience with certain relationships (clues) that will exist between parts of the retinal images of different objects in space. The more important of these visual clues are: (*a*) *Overlay*, by which the images of near objects overlap and tend to hide those of the more distant objects. (*b*) *Perspective*, which depends upon the fact that objects of equal size have smaller retinal images when at a distance than when near by. Linear perspective relates to the apparent convergence of parallel lines that recede in the distance (railroad tracks, etc.). Details within known objects are more readily seen when near than when distant. Thus, the size of retinal image related to known size provides the clue for estimation of distance. (*c*) *Aerial perspective*, through which the edges of objects at a distance are less clearly defined than those near by. Moreover, the more distant objects appear cooler (bluer) in color on account of atmospheric haze. Near objects appear brighter with more color saturation than those more distant. (*d*) *Light and shadow*, which give clues as to shapes and relative positions of objects. (*e*) *Parallax*, which results from head movements, for the relative alignment of more distant objects changes less than that for near objects with the same movement. This clue to depth perception is very strong, and the precision of depth estimation through it is nearly as great as that of stereoscopic depth perception (Tschermak). (*f*) *Height*, whereby objects seen above others are also judged more distant. (*g*) To a small extent accommodation and convergence, through a proprioceptive sense arising from the muscles of the eyes, may provide a clue for gross differences in depth.

As is well known it is impossible to present a single bidimensional picture which will have all the characteristics of the actual three-dimensional scene being portrayed. This is especially true when the picture is being viewed binocularly. As the eyes move over a painting, no change in accommodation or convergence is demanded and obviously no disparity clues are present for stereoscopic space localization of the details in the picture as there would be in the actual scene. The perspective in a pictorial representation of any scene, whether a photograph or a painting (except in certain art styles), refers to a fixed station point and in order to appear correct must be viewed from that point. A contact-print photograph should, then, be viewed at that distance that approximates the focal length of the camera that took the photograph. Viewed at other distances the perspective is exaggerated and unnatural. The viewing distance of an enlarged photograph will be equal to the product of the focal length of the camera and the magnification of the enlargement. This larger viewing distance explains the more pleasing effect derived from looking at enlarged photographs. It is desirable that pictures be viewed in such a way that the observer feels himself a part of the scene. The screen or plane of the picture itself should seem detached from the surroundings and preferably should appear indefinitely localized. To accomplish this is not always easy; usually it can be only approximated, especially for pictures viewed near by with binocular vision. It must be pointed out, however, that with the proper attitude on the part of the observer, and in the absence of distracting peripheral detail, the several psychological constancy phenomena tend to correct small distortions of form, of size, and of the color-brightness relationships.

13. BINOCULAR VISION

Binocular vision is the coordinated use of the two eyes, in which a single perception of external space is obtained, and by which the specific sensation of stereoscopic depth perception is made possible. The final perceptual images from the two eyes are normally said to *fuse* in the brain, and through the strength of the fusion impulse all movements of the eyes become coordinated in the process of fixating different points in space.

One differentiates between the binocular visual field, which will correspond to that portion of visual space for which the images of the two eyes will overlap, and the binocular field of fixation which is the maximum field swept over by the movements of the two eyes. Figures 24 and 25 illustrate the approximate angular dimensions of these fields.

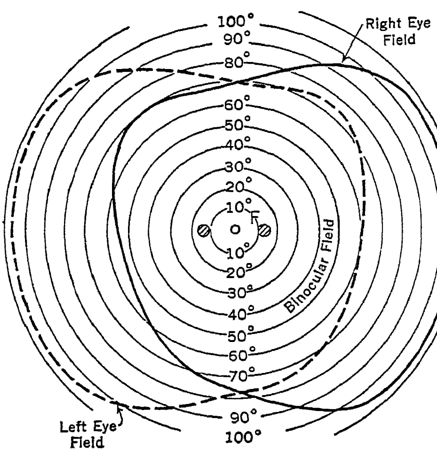


FIG. 24. The Binocular Visual Field (after Southall, *Introduction to Physiological Optics*)

The luminosity of objects is somewhat increased by the use of two eyes over that of only one, depending upon experimental conditions. Also the visual acuity of two eyes is usually found to be higher than that of one eye.

FUSIONAL AREAS. When the eyes are fixated upon one point in space, not all other points in space are perceived single. In fact, only those points will be seen single that are situated within a certain three-dimensional region determined by the distance of the fixation point and by the equivalent anatomical extent of certain areas on the retinas (Panum's fusional areas). See Fig. 25. Points in space outside these areas would ordinarily be seen double (physiological diplopia); however, unless attention is called to them, one is usually suppressed. The functional extent of Panum's area varies to some extent with visual conditions, and its measurement of size decreases somewhat with practice. The size increases away from the foveas toward the periphery. Near the fovea its minimum extent is about 6 to 12 minutes of arc in the horizontal meridian and about 6 minutes of arc in the vertical meridian. All objects beyond a fixation point at about 50 ft from the observer will generally be seen single. Double images exert compulsion innervations for the eyes to move so as to overcome the doubling. In the horizontal meridian, the actual eye movements, however, are usually subject to the will.

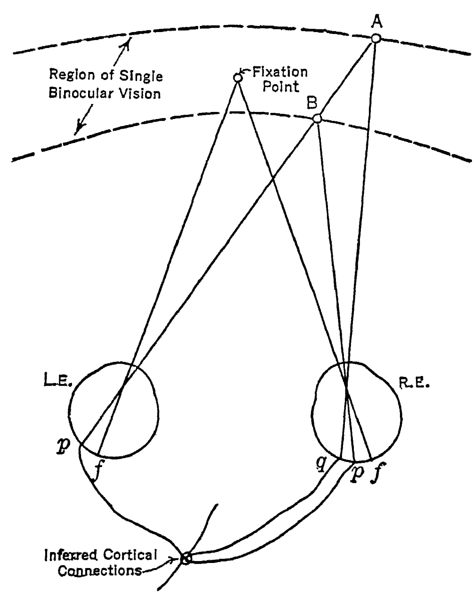


FIG. 25. The Spatial Region of Single Binocular Vision and the Geometrical Relations for the Disparity of Retinal Images in the Two Eyes

STEREOSCOPIC VISION. Stereoscopic vision rests upon the fact that each of the two eyes, by virtue of their separation,* sees objects in the visual field from a slightly different

* The variation of the interpupillary distance, in the general population, is between 55 and 75 mm, with a median at about 63 mm (measured while the eyes are looking at a distant object).

point of view, and hence the two retinal image patterns are slightly dissimilar. See Fig. 26. The angular separation of the retinal images, arising from two points in space, will be different in the two eyes, if one of the points is farther away from the observer. This difference in angular separation defines a *disparity* between the images. A disparity must always relate to two points in space. Referring again to Fig. 25, the double images from points at A and beyond are said to be *uncrossed* disparate with respect to the images of F, for if the right eye is suddenly closed the right half-image vanishes. Similarly the images of points at B and nearer are said to be *crossed* disparate with respect to the images of F, for if the right eye is suddenly closed the left half-image vanishes. Between the points A and B there will be some position for which the images from the two eyes will not be disparate, either crossed or uncrossed with respect to the image of F. This criterion defines the position of points on the horopter surface. Stereoscopic depth perception arises by virtue of the disparity between retinal images, uncrossed disparity being associated with the sense of distance away from the fixated point. Objects seen in crossed disparity are seen nearer than that point. Stereopsis is a specific sensation resting upon

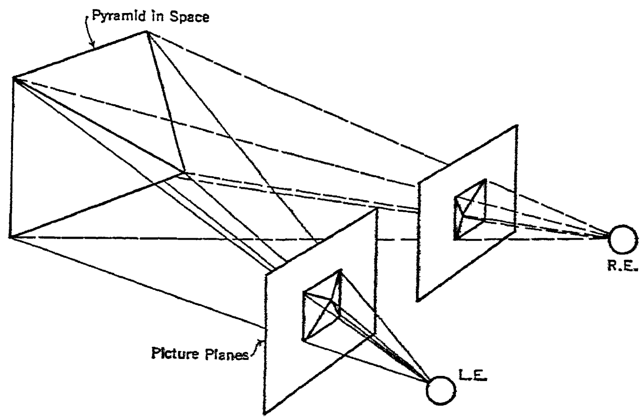


Fig. 26. Illustration of the Difference in the Image Patterns of the Two Eyes in Binocular Space Perception

the physiological and anatomical organization of the retinal elements of the two eyes. It usually occurs immediately for almost instantaneous illumination, and it exists within the entire binocular visual field.

The angular disparity between the images of two separated objects in depth in space will increase with the separation of the eyes and will decrease with their distance from the eyes. Though a sensation of greater depth will be associated with greater disparities of the images, the effect is not rigidly geometrical. For a given angular disparity, the apparent depth interval, to be quantitative, must vary with the square of the visual distance.

The visual acuity of the poorer eye must ultimately limit the threshold of stereoscopic vision. Likewise, the threshold of stereoscopic vision will depend upon those factors influencing visual acuity and, like visual acuity, will vary with the duration of the stimuli. For central vision and for durations longer than 3 sec, maximum acuity is obtained; for shorter durations down to 0.2 sec there is a 4 to 5 fold increase in threshold, after which the instantaneous threshold is nearly constant (Langlands).

Although varying with interpupillary distance and with individuals as well as with the nature of the detail in the visual field, the limiting threshold for stereopsis frequently is taken, on the average, as 30 seconds of arc. There is, correspondingly, a visual distance beyond which objects at greater distances cannot be seen stereoscopically. This limiting distance will roughly be 450 meters, or about a quarter of a mile. Under certain conditions, when the threshold of stereopsis may be as low as 6 seconds of arc, this distance will be exceeded several times. For the peripheral regions of the retina, this limiting distance will be much reduced, owing to the lowered acuity.

Two points separated vertically will have images in the two eyes which are vertically disparate if those points are located to the right or left of the plane perpendicular to the interpupillary base line. Vertical disparities do not give rise to the perception of depth, as do horizontal disparities, but they undoubtedly aid in the spatial localization of objects.

Since the invention of the stereoscope by Wheatstone in 1838, and the discovery of the bases for stereoscopic vision, various instruments have been devised whereby stereoscopic views could be obtained with pictures. In general, the procedure consists in presenting slightly different pictures or drawings before the two eyes which would correspond to the views that each of the eyes would perceive if they had been present when the photograph was taken. The devices include the mirror- and prism-stereoscopes, the haploscope, the red and green and polaroid anaglyphs, grid devices, and even motion pictures where the left- and right-eye views are projected alternately. The problem of projecting pictures for stereoscopic vision, so as to preserve correct disparity relationships and the correct perspective features, is a difficult one, especially for a group of observers (Rule).

RIVALRY. If the fields presented to the two eyes are greatly different, for example in radically different colors or in detail, instead of fusion or even a simultaneous perception of both patterns taking place, *retinal rivalry* occurs. In this, either one field or the other is seen, usually alternately. In the case of dissimilar patterns sometimes sections of each are seen simultaneously, but seldom both in the same region of the visual field. The period of alternation of the two visual fields varies between 2 and 12 sec, depending upon differences in luminosity, area, distinctness of detail within the fields, and central or peripheral vision. One field may prevail over the other for longer periods if there is a great difference in luminosity, or if intelligible detail exists on one and not the other, etc. Ocular dominance may also be a factor in the field that prevails the longer. Only in the case of certain color differences can fusion and therefore the emergence of a mixed color arise.

BIBLIOGRAPHY

General

Helmholtz, H. v., *Physiological Optics*, Eng. trans. by J. P. C. Southall. Optical Soc. Am. (1925).
Duke-Elder, W. S., *Text-Book of Ophthalmology*, Vol. 1. Kimpton, London (1932).
Polyak, S. L. *The Retina*. University of Chicago Press (1941).
Walls, G., *The Vertebrate Eye*. Cranbrook Institute of Science, *Bull.* 19 (August 1942).
Southall, J. P. C., *Introduction to Physiological Optics*. Oxford University Press (1937).
Emsley, H. H., *Visual Optics*, 3d Ed. Hatton Press, London (1944).
Bartley, S. H., *Vision, A Study of Its Basis*. Van Nostrand (1941).
Guilford, J. P., *Psychometric Methods*, Chapters IV–VI. McGraw-Hill (1936).
Luckiesh, M., and F. K. Moss, *The Science of Seeing*. Van Nostrand (1937).
Hardy, A. C., *Handbook of Colorimetry*. Mass. Inst. of Technology (1936).
Moon, P., *The Scientific Basis of Illuminating Engineering*, Chapter XII. McGraw-Hill (1936).
Woodworth, R. S., *Experimental Psychology*, Chapters 22–26. Henry Holt (1938).
Hering, E., Raumsinn, *Hermans Handbuch d. Physiol. d. Sinnesorgane*. An English translation, *Spatial Sense and Movements of the Eye*, C. A. Radde. American Academy of Optometry, Baltimore (1942).

REFERENCES

Asher, L., Monoculares und binoculares Blickfeld eines Emmetropen, *Arch. f. Ophth.*, Vol. 48, 427–431 (1899).
Bond, M. E., and D. Nickerson, Color-order Systems. Munsell and Ostwald, *J. Optical Soc. Am.*, Vol. 32, 709 (December 1942).
Cobb, P. W., The Influence of Pupillary Diameter on Visual Acuity. *Am. J. Physiol.*, Vol. 36, 335 (1915).
Cobb, P. W., and F. K. Moss, The Four Variables of the Visual Threshold, *J. Franklin Inst.*, Vol. 205, 831 (June 1928).
Cobb, P. W., and F. K. Moss, Glare and the Four Fundamental Factors of Vision, *I.E.S. Trans.*, Vol. 23, 1104 (1928).
De Silva, H. R., An Experimental Investigation of the Determinants of Apparent Visual Movement, *Am. J. Psych.*, Vol. 37, 461–501 (1926).
Dohner, D. R., and C. E. Foss, Color-mixing Systems. Color vs. Colorant Mixture, *J. Optical Soc. Am.*, Vol. 32, 702 (December 1942).
Duane, A., Studies in Monocular and Binocular Accommodation, etc., *Am. J. Ophthal.*, 3rd Series, Vol. 5, 865 (1922).
Engstrom, E. W., A Study of Television Image Characteristics, Pt. II, *Proc. I.R.E.*, Vol. 23, 295 (1935).
Freeman, E., Anomalies of Visual Acuity in Relation to Stimulus-distance, *J. Optical Soc. Am.*, Vol. 22, 285 (1932).
Hecht, S., Quantum Relations of Vision, *J. Optical Soc. Am.*, Vol. 32, 42 (January 1942).
Hecht, S., Relation between Visual Acuity and Illumination, *J. Gen. Physiol.*, Vol. 36, 335 (1915).
Hecht, S., The Visual Discrimination of Intensity and the Weber-Fechner Law, *J. Gen. Physiol.*, Vol. 7, 235–267 (1924).
Hecht, S., The Nature of the Photoreceptor Process, Chapter 14, *Murchinson's Handbook of General Experimental Psychology*. Clark Univ. (1934).
Hecht, S., and E. U. Mintz, Visibility of Single Lines, etc., *J. Gen. Physiol.*, Vol. 22, 593 (1939).
Hecht, S., and R. E. Williams, The Visibility of Monochromatic Radiation and the Absorption Spectrum of the Visual Purple, *J. Gen. Physiol.*, Vol. 5, 1–33 (1922).
Jones, L. A., The Fundamental Scale for Pure Hue and Retinal Sensibility to Hue Differences, *J. Optical Soc. Am.*, Vol. 1, 63–77 (1917).
Kelley, K. L., Color Designations for Lights, *J. Optical Soc. Am.*, Vol. 33, 627 (November 1943).
Klein, G. S., Relation between Motion and Form Acuity in Parafoveal and Peripheral Vision and Related Phenomena, *Arch. Psychol.*, Vol. 39, 1–69 (October 1942).

Langlands, N. M. S., Experiments on Binocular Vision, *Trans. Optical Soc. London*, Vol. 28, 45–103 (1927).
MacAdam, D. L., Visual Sensitivities to Color Differences in Daylight, *J. Optical Soc. Am.*, Vol. 32, 247 (May 1942).
Moon, P., and D. E. Spencer, Visual Data Applied to Lighting Design, *J. Optical Soc. Am.*, Vol. 34, 605 (October 1944).
Munsell, A. H., *Munsell Book of Color*. Munsell Color Co., Baltimore (1929).
Niven, J. J., and R. H. Brown, Visual Resolution as a Function of Intensity and Exposure Time of the Human Fovea, *J. Optical Soc. Am.*, Vol. 34, 738 (December 1944).
Priest, I. G., and F. G. Brickwedde, cf. S. Hecht, Interrelations of Various Aspects of Color Vision, *J. Optical Soc. Am.*, Vol. 21, 615–639 (1931).
Reeves, P., Response of the Pupil to Various Intensities of Light, *J. Optical Soc. Am.*, Vol. 4, 35 (1920).
Rowland, W. M., and L. L. Sloan, The Relative Merits of Red and White Light of Low Intensity for Adapting the Eyes to Darkness, *J. Optical Soc. Am.*, Vol. 34, 601 (October 1944).
Rule, J. R., Stereoscopic Drawings, *J. Optical Soc. Am.*, Vol. 28, 313–322 (1938); also The Shape of Stereoscopic Images, *ibid.*, 31, 124–129 (1941).
Stiles, W. S., and B. H. Crawford, Luminous Efficiency of Rays Entering the Eye Pupil at Different Points, *Proc. Roy. Soc.*, Vol. 112B, 428 (1933).
Wald, G., Human Vision and the Spectrum, *Science*, Vol. 101, 653 (June 29, 1945).
Walls, G., Factors in Human Visual Resolution, *J. Optical Soc. Am.*, Vol. 33, 489 (September 1943).
Wertheim, T., Indirect Visual Acuity, *Zeitschrift für Psych. und Physiol. d. Sinnesorg.*, Vol. 7, 172 (1894).

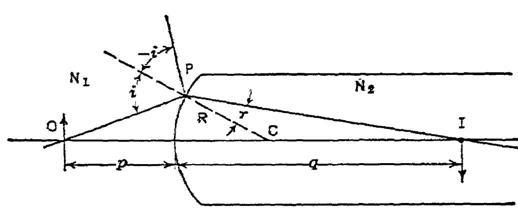
ELECTRON OPTICS

By D. W. Epstein

GENERAL CONCEPTS. Electron optics has become a branch of applied physics. The term "electron optics" was chosen because of the similarity between the path of an electron, or any charged particle, moving through electrostatic and magnetostatic fields and that of a ray of light passing through refracting media. Because of this similarity, such concepts of optics as lens and focal length may be transferred to electrostatic and magnetostatic fields.

The subject of optics is generally divided into: (1) geometrical optics, which treats only of the geometrical relations of the propagation of light; and (2) physical optics, which, utilizing the wave theory of light, is capable of dealing with any problem in light. A similar division is made in electron optics. This chapter will concern itself exclusively with non-relativistic geometrical electron optics of static fields; that is, relativity, wave mechanics, and electron optics of high-frequency phenomena will be excluded.

The analogy between electron optics and light is indicated in Figs. 1 and 2. Figure 1 shows the trajectory of a ray of light refracted and reflected at a spherical interface separating two regions of different indices of refraction. The analogous electron optical case is shown in Fig. 2. Assume that by some means (such as two closely spaced meshes at different electrostatic potentials) a space in vacuum is divided into two regions by a sphericocylindrical surface as shown in Fig. 2, the region to the left of the spherical surface being at the electrostatic potential V_1 and the region to the right of the spherical surface being at the electrostatic potential V_2. An electron emitted with zero initial velocity by a cathode (located to the left of the V_1 region) at zero potential will move in the V_1 region at a constant velocity v_1 given by



$$N_1 \sin i = N_2 \sin r$$

$$\frac{N_2}{q} - \frac{N_1}{p} = \frac{N_2 - N_1}{R} = -\frac{N_1}{f_1} = \frac{N_2}{f_2}$$

$$\frac{f_2}{f_1} = -\frac{N_2}{N_1}$$

$$m = \frac{N_1}{N_2}\frac{q}{p}$$

FIG. 1. Trajectory of a Ray of Light Refracted and Reflected at a Spherical Interface Separating Two Regions of Different Indices of Refraction

$$v_1 = \sqrt{2\frac{e}{m}V_1} = 5.95 \times 10^5 \sqrt{V_1} \text{ meters per sec} \tag{1}$$

where $e = 1.59 \times 10^{-19}$ coulomb is the charge of the electron and $m = 9.04 \times 10^{-31}$ kg is its mass. As long as the electron stays in the region of constant potential V_1 there is no electrostatic force acting on it and it will move in a straight line, say OP in Fig. 2; this corresponds to the law of rectilinear propagation in light optics.

When the electron crosses the spherical surface at P its velocity is changed to v_2 corresponding to V_2, i.e., to $v_2 = 5.95 \times 10^5 \sqrt{V_2}$ meters per second. The force being normal to the surface, only the component of velocity v_R normal to the surface will change; the tangential component of velocity v_T will be the same on both sides of the surface. It thus follows (see Fig. 2) that

$$v_T = v_1 \sin i = v_2 \sin r \tag{2}$$

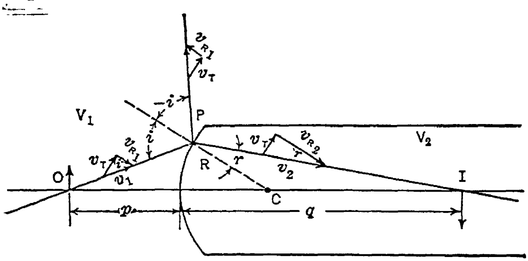
where i and r are the angles of incidence and refraction, respectively. Equation (2) may also be written

$$\sqrt{V_1} \sin i = \sqrt{V_2} \sin r$$

or

$$\frac{\sin i}{\sin r} = \frac{v_2}{v_1} = \frac{\sqrt{V_2}}{\sqrt{V_1}} = \frac{N_2}{N_1} = N \tag{3}$$

so if N_1 and N_2 are identified as the indices of refraction of the left region and right region respectively and N as the relative index of refraction then eq. (3) becomes the well-known law of refraction. The focusing properties of a spherical refracting surface then follow as indicated in Fig. 2.



$$v_1 = \sqrt{2\tfrac{e}{m} V_1} \qquad v_2 = \sqrt{2\tfrac{e}{m} V_2}$$

$$v_T = v_1 \sin i = v_2 \sin r$$

$$\sqrt{V_1} \sin i = \sqrt{V_2} \sin r$$

$$\frac{\sqrt{V_2}}{q} - \frac{\sqrt{V_1}}{p} = \frac{\sqrt{V_2} - \sqrt{V_1}}{R} = -\frac{\sqrt{V_1}}{f_1} = \frac{\sqrt{V_2}}{f_2}$$

$$\frac{f_2}{f_1} = -\sqrt{\frac{V_2}{V_1}}$$

$$m = \sqrt{\frac{V_1}{V_2}}\ \frac{q}{p}$$

$$V_2 < V_1$$

$$|V_2 - V_1| > \frac{1}{2} m (v_1 \cos i)^2$$

FIG. 2. Electron Optical Analogy of Fig. 1

If $V_2 < V_1$, then $e(V_2 - V_1)$ is negative, and if in absolute magnitude it is greater than $1/2\ m\ (v_1 \cos i)^2$—the part of kinetic energy of the electron corresponding to the normal component of its velocity—then the electron will be shot back from the surface with its normal velocity component reversed. The path of the reflected electron will make with the normal to the surface the same angle i which its path made on incidence (see Fig. 2). Thus the law of reflection also holds in electron optics.

The different rays of a beam of light do not affect one another. The various electrons in an electron beam repel one another; this, of course, is the well-known effect of space charge. However, for low electron beam intensities the effect of space charge is negligible, and it will be so assumed in what follows.

It is generally customary to deduce the laws of geometrical optics especially for nonhomogeneous media from Fermat's principle of shortest optical path.

This principle states that the path of a ray of light from a point A to a point B is always such as to make the integral an extremum (usually a minimum) with respect to all neighboring paths for rays of the same frequency. The principle is usually stated as

$$\delta \int_A^B N\, ds = 0 \tag{4}$$

where N, the index of refraction, may be a function of direction as well as position and ds is an element of path length.

In particle dynamics there is the similar principle of least action stating that

$$\delta \int_A^B p\, ds = 0 \tag{5}$$

where p is the momentum of the particle. A comparison of eqs. (4) and (5) shows that an electron in an electrostatic and magnetostatic field will follow the same trajectory as light would if the index of refraction at every point were made proportional to the momentum of the electron at the point. The momentum of an electron moving in an electrostatic and magnetostatic field is

$$p = mv - eA \cos \theta$$

where θ is the angle between the direction of motion of the electron and the direction of magnetic vector potential A. The index of refraction for an electron moving in a combined electrostatic and magnetostatic field is

$$N = k\left[v - \frac{e}{m} A \cos \theta\right] \tag{6}$$

Equation (6) shows that owing to the magnetostatic field N is a function not only of the position of the electron but also of its direction of motion so that combined electrostatic and magnetostatic fields behave like non-homogeneous anisotropic media.

It was noted in Fig. 2 that the index of refraction for an electron at a point in an electrostatic field is proportional to its speed at the point, i.e., $N = kv$, but it must not be assumed that the index of refraction in a magnetostatic field is $ke/mA \cos \theta$ but rather

$$k\left[v_0 - \frac{e}{m} A \cos \theta\right] \tag{7}$$

where v_0 is the constant speed with which the electron moves through the magnetic field.

14. ELECTROSTATIC LENSES

The electrostatic lens of Fig. 2 is an example of a relatively impractical case. Practical electrostatic lenses are generally formed by the application of electrostatic potentials to axial symmetric electrodes. Figure 3 shows a cross-section through the axis of such an electron focusing system. The double lines represent cylindrical, hollow conductors at the potentials V_1 and V_2; the single lines represent the equipotential surfaces in the space (vacuum) between the electrodes. From eq. (3) it follows that each equipotential surface represents a surface of constant index of refraction. Here are shown only a few of the equipotential surfaces; actually, of course, there is an infinite series of equipotential surfaces having a common axis. This electron focusing system may, therefore, be considered as a very large number of coaxial refracting surfaces. Most optical systems for light consist of a series of spherical refracting surfaces having a common axis of symmetry called the optic axis. For light, the optical systems are usually such that the index of refraction changes abruptly as light passes from one medium to the other. In electron optics, the index of refraction is a continuous function of position. Optically speaking, this means that an electrostatic field constitutes an isotropic non-homogeneous medium for electrons, corresponding to a medium of continuously variable density for light rays.

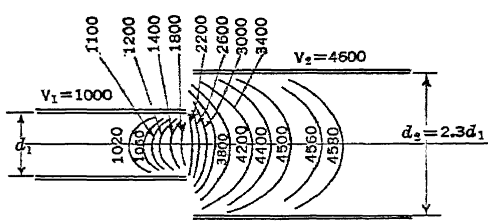


FIG. 3. Equipotential Line Plot in Charge-free Space Due to Potentials Applied to Coaxial Cylindrical Electrodes

The potential distribution $V(r, z)$—the potential at a radial distance r from the axis and an axial distance z from the origin—in charge-free space due to potentials applied to axial symmetric electrodes is given in cylindrical coordinates by the Laplace equation

$$\frac{\partial^2 V(r, z)}{\partial r^2} + \frac{1}{r}\frac{\partial V(r, z)}{\partial r} + \frac{\partial^2 V(r, z)}{\partial z^2} = 0$$

subject to the boundary conditions that $V(r, z)$ at the electrodes assumes the values of the potentials applied to the electrodes. Except for some special cases it has not been possible to obtain mathematical solutions subject to the actually existing boundary conditions. The solutions are generally obtained experimentally by means of an electrolytic tank. The potential distribution of Fig. 3 was thus determined.

The equations of motion of an electron moving in a meridian plane, i.e., a plane containing the axis, are

$$\begin{aligned} m\frac{d^2 z}{dt^2} &= e\frac{\partial V(r, z)}{\partial z} \\ m\frac{d^2 r}{dt^2} &= e\frac{\partial V(r, z)}{\partial r} \end{aligned} \tag{8}$$

The trajectory of a meridional electron traversing an axially symmetric electrostatic field $V(r, z)$ and with a velocity $v = \sqrt{2(e/m)V(r, z)}$, namely, an electron emitted with zero velocity from a cathode at zero potential, is given by the differential equation

$$\frac{d^2r}{dz} + \frac{\left[1 + \left(\frac{dr}{dz}\right)^2\right]}{2V(r, z)} \frac{\partial V(r, z)}{\partial z} \frac{dr}{dz} - \frac{\left[1 + \left(\frac{dr}{dz}\right)^2\right]}{2V(r, z)} \frac{\partial V(r, z)}{\partial r} = 0 \tag{9}$$

The distribution of potential in space is uniquely determined if the distribution along the axis together with its even derivatives are known. Thus

$$V(r, z) = V(0, z) - \frac{r^2}{2^2} V''(0, z) + \frac{r^4}{2^2 4^2} V^{\mathrm{IV}}(0, z) + \cdots \tag{10}$$

where $V(0, z)$ is the distribution of potential along the axis, $V''(0, z)$ is the second derivative of axial potential with respect to z, $V^{\mathrm{IV}}(0, z)$ is the fourth derivative, etc.

It may be shown by using eq. (10) that the equipotential surfaces in the neighborhood of the axis are hyperboloids with a radius of curvature of

$$R = \frac{2V'(0, z)}{V''(0, z)} \tag{11}$$

This shows that, in electron optics, index of refraction and curvature cannot be varied independently as they can in light optics. Consequently, the correction of some lens aberrations is extremely difficult if not impossible in electron optics.

An optical system is usually described in terms of paraxial or first-order imagery. Actual imagery departs from paraxial imagery. Such departures are described as aberrations.

The focusing action of an electrostatic field is similarly described to a first approximation by considering only paraxial electrons. Paraxial electrons are characterized by the fact that in calculating their paths it is assumed that their distances from the axis, r, and their inclination toward the axis, dr/dz, are so small that the second and higher powers of r and dr/dz are negligible.

The differential equation for the trajectory traversed by a paraxial electron becomes from eq. (9)

$$\frac{d^2r}{dz^2} + \frac{V'}{2V}\frac{dr}{dz} + \frac{V''}{4V} r = 0 \tag{12}$$

or

$$\sqrt{V}\frac{d}{dz}\left(\sqrt{V}\frac{dr}{dz}\right) + \frac{1}{4} V'' r = 0 \tag{13}$$

where V, V', and V'' are the axial distribution of potential $V(0, z)$ and its first two derivatives with respect to z respectively. Equation (12) or its equivalent eq. (13) may be taken as the fundamental equation of paraxial electron optics of axially symmetric electrostatic fields.

The general solution of the second-order differential equation (12) or (13) may be written

$$r(z) = c_1 r_1(z) + c_2 r_2(z) \tag{14}$$

where $r_1(z)$ and $r_2(z)$ are any two linearly independent solutions and c_1 and c_2 are arbitrary constants. Equation (14) states that the trajectory of any paraxial electron is simply the linear combination of two independent trajectories. Hence the complete paraxial focusing action of an axially symmetric electrostatic field is determined by calculating the trajectories of only two electrons. The trajectories of two electrons entering the lens parallel to the axis from the object and image sides are chosen as the two solutions $r_1(z)$ and $r_2(z)$ and are called the two *fundamental trajectories*. These trajectories determine the location of the cardinal points of the focusing system, i.e., the location of the focal and principal points, and thus the focusing action of the lens.

Referring to Fig. 4 let S_1 and S_2 be two equipotential surfaces such that the space to the left of S_1 is equipotential and is at potential V_1 and the space to the right of S_2 is equipotential and at the potential V_2. The potential in the region between S_1 and S_2 varies continuously, as shown in Fig. 3. Then the paraxial electron moving parallel to the axis in the equipotential space to the left of S_1 will follow the trajectory $r_1(z)$ and after passing through the focusing system will move in a direction inclined at an angle to the axis and will pass through the axial point F_2. All paraxial electrons moving parallel to the axis in the V_1 or object space will pass through F_2. The point F_2 is the second focal point. The plane passing through the second focal point and perpendicular to the axis of symmetry is the second focal plane.

The plane perpendicular to the axis and passing through the point of intersection of the original and final directions of motion of the electron is the second principal plane. The point of intersection H_2 between the second principal plane and the axis is the second principal point. The distance H_2F_2 denoted by f_2 is the second focal length.

Similarly, a paraxial electron moving parallel to the axis in the V_2 or image space will after passing through the focusing system move in a direction inclined to the axis and will

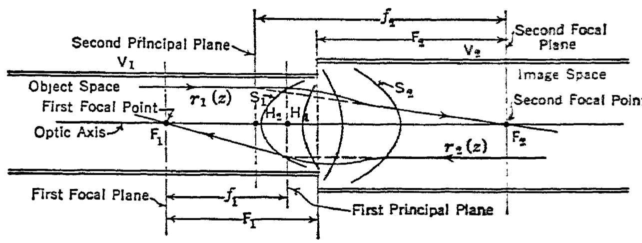


FIG. 4. Fundamental Trajectories and Cardinal Points of an Electrostatic Lens

pass through the point F_1 in the object space. F_1 is the first focal point. F_1 may also be considered that axial point in the object space from which all electrons, after passing through the focusing system, are parallel to the axis in the image space.

The plane perpendicular to the axis of symmetry and passing through the first focal point is the first focal plane. The plane perpendicular to the axis and passing through the point of intersection of the original and final directions of motion of the electron is the first principal plane. The point of intersection, H_1, of the first principal plane and the axis is the first principal point.

It is to be noted that in Fig. 4 the principal planes are crossed. This is a characteristic of lenses having indices of refraction different on the two sides.

In Fig. 5 let A_1B_1 be an object (say, an aperture through which electrons are passing) located in the equipotential region V_1. Then a paraxial electron coming from A_1 and moving parallel to the optic axis will after passing through the lens go in the direction F_2A_2. A paraxial electron issuing from A_1 in the direction A_1F_1 will after passing through the lens move parallel to the optic axis and will intersect the trajectory of the other electron at A_2. Similarly the trajectories of all paraxial electrons coming from A_1 will intersect at A_2, the image of A_1. The same is true of every point on A_1B_1, and so the inverted image

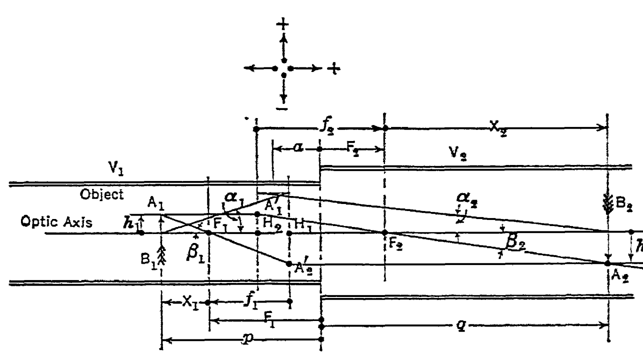


FIG. 5. Image Formation in a Direct Bipotential Electrostatic Lens

A_2B_2 is obtained. The ratio A_2B_2/A_1B_1 or h_2/h_1 gives the magnification. The electron image of A_1B_1 becomes visible if a fluorescent screen is placed in the plane of A_2B_2.

From eq. (12) or (13) and Fig. 5 it may be shown that the following relations hold:

$$\frac{f_2}{f_1} = -\sqrt{\frac{V_2}{V_1}} \tag{15}$$

$$X_1X_2 = f_1f_2 \tag{16}$$

$$\frac{f_1}{f_1 + X_1} + \frac{f_2}{f_2 + X_2} = \frac{f_1}{p + (f_1 - F_1)} + \frac{f_2}{q + (f_2 - F_2)} = 1 \tag{17}$$

$$\frac{1}{[q + (f_2 - F_2)]} - \frac{1}{\sqrt{\frac{V_2}{V_1}}\,[p + (f_1 - F_1)]} = \frac{1}{f_2} \tag{18}$$

$$m = \frac{h_2}{h_1} = -\frac{f_1}{X_1} = -\frac{X_2}{f_2} = \sqrt{\frac{V_1}{V_2}}\frac{[q + (f_2 - F_2)]}{[p + (f_1 - F_1)]} = \sqrt{\frac{V_1}{V_2}}\frac{\beta_1}{\beta_2} = \sqrt{\frac{V_1}{V_2}}\frac{\alpha_1}{\alpha_2} \tag{19}$$

$$p = (F_1 - f_1) + f_1\frac{(m - 1)}{m} \tag{20}$$

$$q = (F_2 - f_2) - f_2(m - 1) \tag{21}$$

The complete paraxial focusing action of an electrostatic lens is determined by means of the above relations when the positions of the cardinal points or f_1, f_2, F_1, and F_2 are known.

Electrostatic lenses are classified according to (*a*) electrode symmetry, (*b*) thickness, and (*c*) the potentials on the sides of the lens.

Spherical lenses are formed by applying different voltages to two or more electrodes having axial symmetry such as apertures and cylindrical and conical tubes. Spherical lenses are used in cathode-ray tubes, television pick-up tubes, and electron microscopes.

Cylindrical lenses are formed by applying different voltages to two or more pairs of electrodes having a plane of symmetry such as pairs of wires or pairs of strips. Cylindrical lenses are used in some receiving and transmitting tubes.

Thick lenses are characterized by the fact that the axial extension of the electrostatic or refracting field is of the same order of magnitude as or even larger than the focal length. It is necessary to know the positions of all four cardinal points in order to determine the focusing action of a "thick lens."

Thin lenses are characterized by the fact that the axial extension of the refracting field is negligible compared with the focal length. In a thin lens the principal planes coincide, and the focusing action of the lens is determined by its location and focal lengths. For estimating purposes it is sufficiently accurate to consider most lenses "thin."

The focal lengths of a "thin" spherical lens may be calculated from the relations

$$\begin{aligned} \frac{1}{f_2} &= \frac{3}{16}\left(\frac{V_1}{V}\right)^{1/4}\int_a^b \left(\frac{V'}{V}\right)^2 dz \\ \frac{1}{f_1} &= -\frac{3}{16}\left(\frac{V_2}{V_1}\right)^{1/4}\int_a^b \left(\frac{V'}{V}\right)^2 dz \end{aligned} \tag{22}$$

where V_1 and V_2 are the potentials of the equipotential regions on the two sides of the lens, V and V' the axial distribution of potential and its first derivative with respect to z, and the integral is taken over the axial extension of the field.

Equations (22) also apply to "thin" cylindrical lenses, if the numerical factor $^3/_{16}$ is replaced by $^1/_2$.

Unipotential lenses are characterized by having identical equipotential regions on the object and image sides, $V_1 = V_2$, and hence $f_1 = -f_2$. Figure 6 shows a few electrode arrangements and axial distributions of potential of unipotential lenses. Figures 7 and 8 * give the focal length of several thin unipotential lenses as a function of V_c/V_0. The dotted curves of Fig. 8 give the focal length of a unipotential lens where the central aperture has been replaced by a fine metal screen or an electron-permeable conducting membrane. It should be noted that this lens is divergent, i.e., f_2 is negative, when V_c/V_0 is less than unity. In the case of thin unipotential lenses eqs. (15) to (21) simplify to

$$f_1 = -f_2 \tag{15'}$$

$$X_1X_2 = -f_2^2 \tag{16'}$$

$$\frac{1}{q} - \frac{1}{p} = \frac{1}{f_2} \tag{18'}$$

$$m = \frac{h_2}{h_1} = \frac{f_2}{X_1} = -\frac{X_2}{f_2} = \frac{q}{p} = \frac{\beta_1}{\beta_2} = \frac{\alpha_1}{\alpha_2} \tag{19'}$$

$$p = -f_2\left(\frac{m - 1}{m}\right) \tag{20'}$$

$$q = -f_2(m - 1) \tag{21'}$$

* Figures 7 and 8 were plotted from data calculated by Dr. E. G. Ramberg of RCA Laboratories.

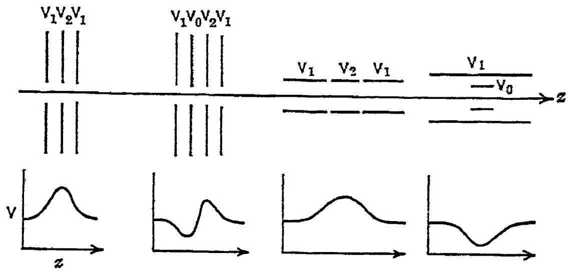


FIG. 6. Electrode Arrangements and Axial Distribution of Potential of Some Unipotential Lenses

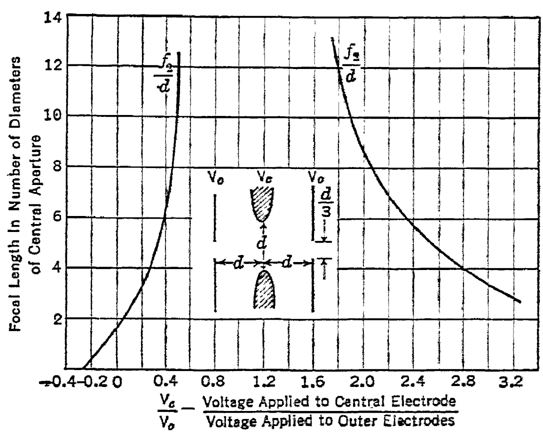


FIG. 7. Variation of the Focal Length (Measured as the Number of Diameters of the Central Aperture) with the Ratio of the Voltages Applied to the Central and Outer Electrodes

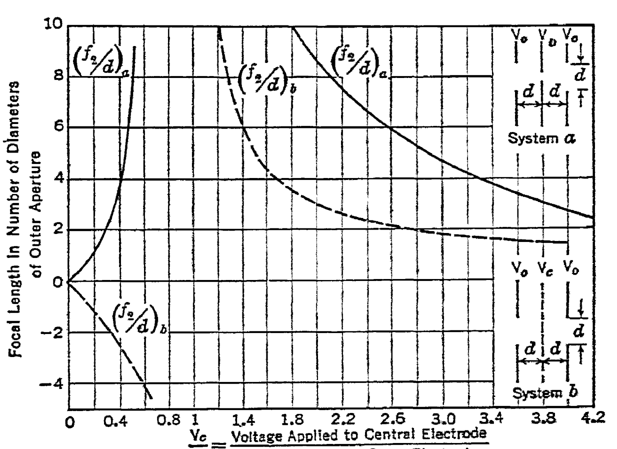


FIG. 8. Variation of the Focal Length (Measured as the Number of Diameters of the Outer Electrodes) with the Ratio of the Voltages Applied to the Central and Outer Electrodes. The solid curves apply to system (*a*) with a small central aperture. The dotted curves apply to system (*b*) with a central electrode consisting of a fine metal mesh or an electron-permeable conducting membrane.

Bipotential lenses are characterized by having different equipotential regions on the object and image sides, or $V_1 \neq V_2$. In the *direct bipotential* lens the potential of the image space is greater than that of the object space ($V_2 > V_1$). In the inverted bipotential lens $V_2 < V_1$. Figure 9 shows a few electrode arrangements and axial distributions of potential of some bipotential lenses. Figures 10*a* to 10*e* give the focal lengths and positions of

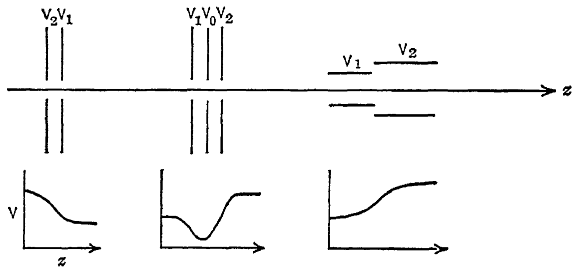


FIG. 9. Electrode Arrangements and Axial Distribution of Potential of Some Bipotential Lenses

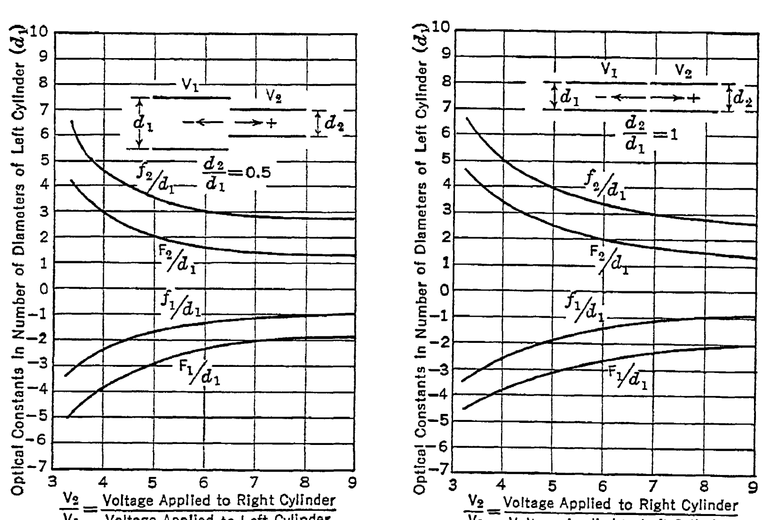
the focal points (thus giving the positions of the four cardinal points, see Fig. 5) as a function of V_2/V_1 for 5 different diameter ratios d_2/d_1. Figure 11 shows f_1, f_2, F_1, and F_2 as a function of d_2/d_1 for $V_2/V_1 = 5$.

The focusing characteristics of a "thick" bipotential lens are given by eqs. (15) to (21). Their use in conjunction with Fig. 10*c* will be illustrated by the following simple example (see Fig. 5 for sign convention). Given the lens with $d_2/d_1 = 1.5$, $V_2/V_1 = 5$, then, from Fig. 10*c*, $f_1 = -2.1d_1$, $(F_1 - f_1) = -1.1d_1$, $f_2 = 4.7d_1$, and $(F_2 - f_2) = -1.8d_1$. It is desired to obtain a real image of an object with a magnification $m = -5$ (since a real image of an object is inverted). Then by eqs. (20) and (21)

$$p = (F_1 - f_1) + f_1\left(\frac{m-1}{m}\right) = -1.1d_1 - 2.1d_1\left(\frac{-5-1}{-5}\right) = -3.6d_1$$

$$q = (F_2 - f_2) - f_2(m-1) = -1.8d_1 + 4.7d_1(-5-1) = 26.4d_1$$

and thus the object is located at a distance of $3.6d_1$ to the left of the cylinder ends and the image $26.4d_1$ to the right of the cylinder ends.



(*a*) (*b*)

FIG. 10. (See facing page.)

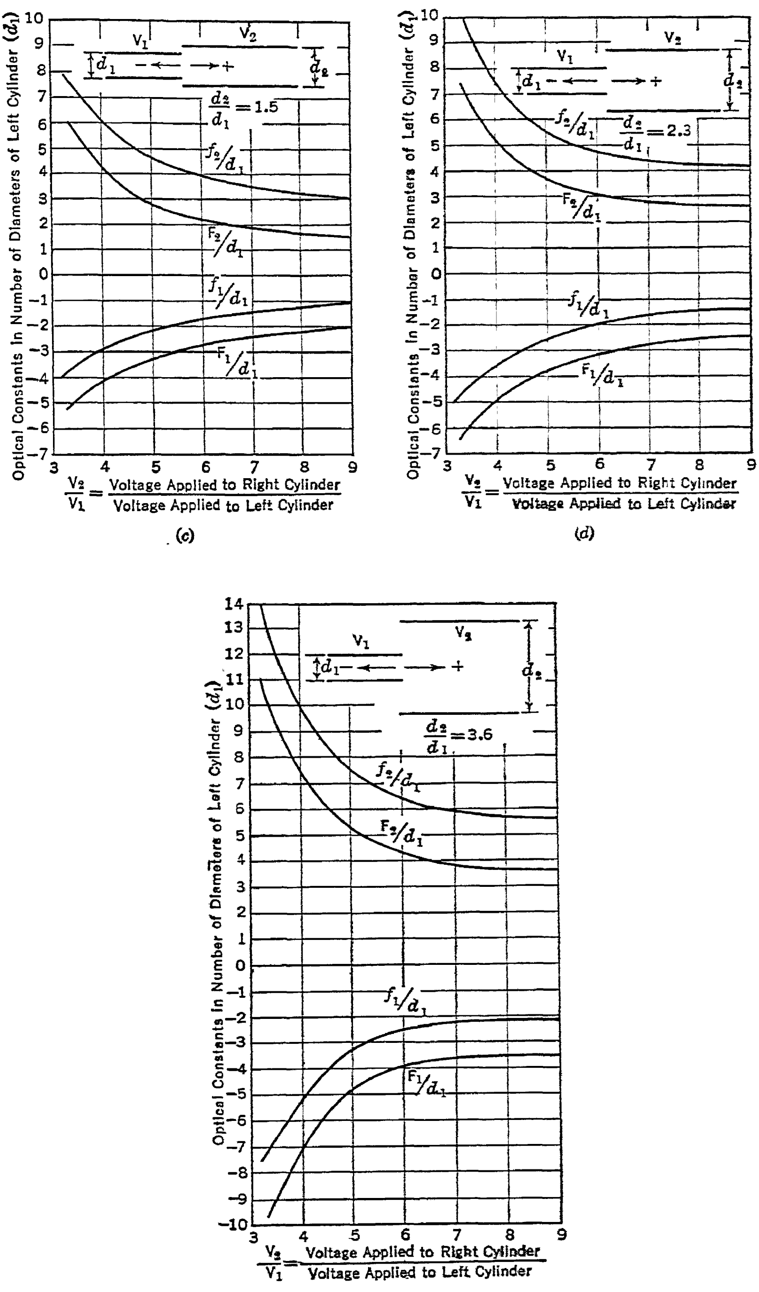


FIG. 10. Variation of Focal Lengths, f, and Distance of Focal Points from End of Cylinder, F (Measured as the Number of Diameters of the Cylinder on the Left), with the Ratio of the Voltages Applied to Right and Left Cylinders

In the case of a "thin" bipotential lens eqs. (18) to (21) simplify to

$$\frac{1}{q + a} - \frac{1}{\sqrt{(V_2/V_1)}(p + a)} = \frac{1}{f_2} \tag{18''}$$

$$m = \frac{h_2}{h_1} = \sqrt{\frac{V_1}{V_2}}\frac{(q + a)}{(p + a)} \tag{19''}$$

$$p = -a + f_1\left(\frac{m - 1}{m}\right) \tag{20''}$$

$$q = -a - f_2(m - 1) \tag{21''}$$

where a is the distance between the position of the equivalent thin lens and cylinder ends as indicated in Fig. 5.

Bipotential lenses are most generally used in cathode-ray tubes.

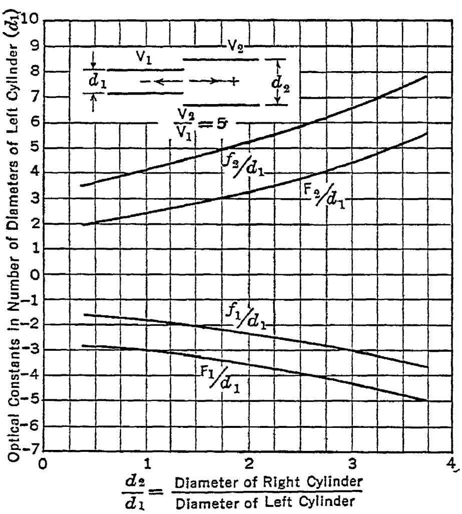


FIG. 11. Variation of Focal Lengths, f, and Distances of Focal Points from End of Cylinders, F (Measured as the Number of Diameters of Cylinder at Left), with the Ratio of Diameters of the Right and Left Cylinders

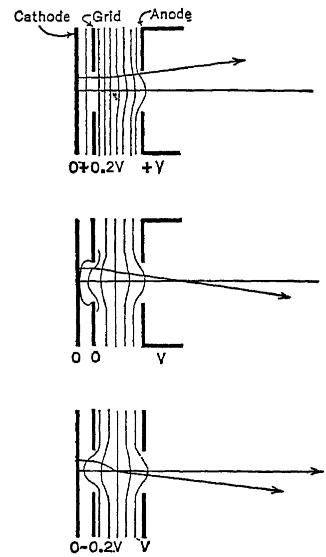


FIG. 12. Electrode Arrangement and Approximate Distributions of Potential (for Three Values of Grid Bias) of One of the Most Commonly Used Cathode Lenses

Cathode lenses are characterized by having a plane of zero potential (cathode) normal to the optic axis. The cardinal points of a cathode lens are of less significance than in case of the other lenses because of the distribution of initial velocities of the electrons emitted by the cathode. Figure 12 shows the electrode arrangement and approximate distribution of potential of one of the most commonly used cathode lenses. Figure 12 also shows the approximate trajectory of an electron emitted normal to the cathode for grid voltages of +0.2, 0, and −0.2 V. The grid is so named since it is also used for controlling the current going to the anode aperture. The cathode lens is the most widely used since it exists in all types of electronic devices.

Electrostatic lenses suffer from the defects or aberrations known as spherical aberration, coma, astigmatism, curvature of field, distortion, and chromatic aberration. They also suffer from defects due to misalignment and malconstruction of electrodes, space charge, relativity effect, etc.

Spherical aberration is in general the most troublesome, especially in instruments requiring a fine focused spot (or line) as in cathode-ray tubes, television pick-up tubes, and electron microscopes. Figure 13 shows the increase in spot size caused by the spherical aberration of a bipotential lens. The increase in spot size is given as a function of the beam diameter at the end of the first cylinder measured in terms of the first-cylinder diameter. Figure 14 shows the decrease in voltage ratio required to focus a beam of diameter d/d_1. In general the spherical aberration of the unipotential lens is greater than that of the bipotential lens and the equidiameter bipotential lens has less spherical aberration than a lens with $(d_2/d_1) > 1$ or $(d_2/d_1) < 1$. In general the aberrations in electron lenses are much more severe than they are in light lenses (this also applies to magnetostatic lenses). This is primarily due to the connection between radius of curvature and index of refraction—see eq. (11)—which exists in electron optics and not in light optics.

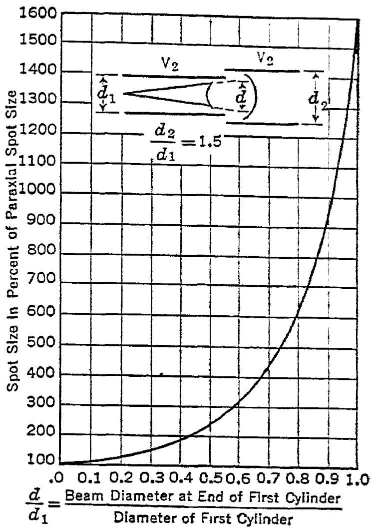


FIG. 13. Increase in Spot Size Caused by the Spherical Aberration of a Bipotential Lens as a Function of the Beam Diameter at the End of the First Cylinder

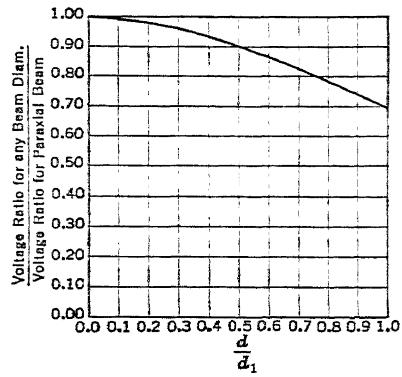


FIG. 14. Decrease in Focusing Voltage Ratio Caused by the Spherical Aberration of a Bipotential Lens as a Function of the Beam Diameter at the End of the First Cylinder

15. MAGNETOSTATIC LENSES

Any axially symmetric magnetostatic field acts as a "lens" for electrons. Such fields are generally created by passing direct current through coils with axial symmetry.

As already noted, a magnetostatic field is an electron-refracting medium whose index of refraction is a function not only of position but also of the *direction* of travel of the electron. Consequently magnetostatic lenses differ in their behavior from electrostatic and ordinary light lenses. In general the differences are made evident by a rotation of the image and the irreversibility of object and image. Thus, a real image, produced by a magnetostatic lens, is not inverted (as it is in an electrostatic lens), but is rotated through an angle θ relative to the inverted image. Similarly, if the image is made the object, then its image will not coincide with the original object (as it would in an electrostatic lens) but will be rotated through an angle 2θ relative to the original object. Magnetostatic focusing is generally accomplished by long or short lenses. The image rotation of a magnetosattic lens is, for a given focal length, reduced as the extent of the field along the axis is reduced. Thus the image rotation is practically zero for an extremely short lens and is 180° (image erect) for a very long lens.

The long lens generally extends over the entire length of the electron beam and is formed by the uniform field of a long solenoid. It is used in such television pick-up tubes as the image dissector, orthicon, and image orthicon. The uniform field of the long lens produces a sequence of uniformly spaced, real, erect (rotation 180°) images with unity magnification of an object which is placed normal to the field and which emits electrons of uniform speeds. This follows from the fact that an electron injected into a uniform magnetostatic field, B, with a speed v, and a corresponding voltage V and at an angle α with the field, describes a helix, the radius of the helix being

$$r = \frac{v \sin \alpha}{(e/m)B} = \frac{3.38 \times 10^{-6}\sqrt{V}}{B} \sin \alpha \text{ meters} \tag{23}$$

$$= \frac{2.69\sqrt{V} \sin \alpha}{H} \text{ meters}$$

The time taken by the electron in describing 1 revolution is

$$T = \frac{2\pi}{(e/m)B} = \frac{3.56 \times 10^{-11}}{B} = \frac{2.84 \times 10^{-5}}{H} \text{ sec} \tag{24}$$

which depends exclusively on B or H and is independent of V or α. The pitch of the helix or the distance that the electron has traveled in time T is

$$s = \frac{2\pi v \cos\alpha}{(e/m)B} = \frac{21.1 \times 10^{-6}\sqrt{V}\cos\alpha}{B} \text{ meters}$$

$$s = \frac{16.8\sqrt{V}\cos\alpha}{H} \text{ meters} \tag{25}$$

If the angles α at which electrons from a point source are injected into the field are small, then their speeds along the lines of force, $v \cos\alpha$, are essentially constant ($\cos\alpha = 1$) and all electrons will reunite at the following distances from the source

$$s_n = \frac{21.1 \times 10^{-6} n\sqrt{V}}{B} = \frac{16.8n\sqrt{V}}{H} \text{ meters} \tag{26}$$

where n is an integer. Since any line of force is an axis of symmetry in a uniform field, an extended electron-emitting (or "illuminated") source placed perpendicular to the field will be imaged at the distances s_n of eq. (26) (see Fig. 15).

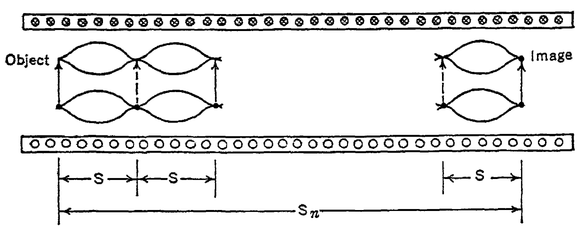


FIG. 15. Image Formation by a Uniform Magnetostatic Field

The short magnetostatic lens extends over a limited region of the beam and is generally formed by the non-uniform field of an iron-encased coil with a narrow slit in the iron shell or pole pieces (Fig. 16). Short lenses are used in electron microscopes and television

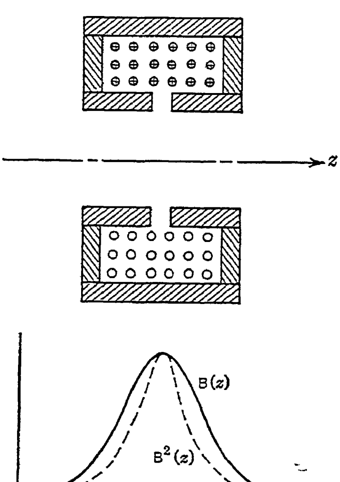


FIG. 16. Short Iron-clad Magnetostatic Lens and Approximate Distribution of Axial Flux Density

cathode-ray tubes. Except for some image rotation the short lens behaves like a thin unipotential electrostatic lens so that its focusing characteristics are given by eqs. (15') to (21′). The focal length of a short lens is given by

$$\frac{1}{f} = \frac{1}{8}\frac{e}{m}\frac{1}{V}\int_a^b B^2\,dz = \frac{2.2 \times 10^{10}}{V}\int_a^b B^2\,dz \quad \text{meters}^{-1}$$
$$= \frac{3.5 \times 10^{-2}}{V}\int_a^b H^2\,dz \quad \text{meters}^{-1} \tag{27}$$

where B is the axial component of the flux density along the axis of the lens, and V is the potential of the equipotential region in which the lens is located.

The image rotation is given by

$$\theta = \sqrt{\frac{1}{8}\frac{e}{m}\frac{1}{v}}\int_a^b B\,dz = \frac{1.5 \times 10^5}{\sqrt{V}}\int_a^b B\,dz \quad \text{radians}$$
$$= \frac{0.19}{\sqrt{V}}\int_a^b H\,dz \quad \text{radians} \tag{28}$$

Integrating eqs. (27) and (28) for a wire loop of diameter d and carrying a current NI results in

$$\frac{f_2}{d} = \frac{8 \times 10^{14} V}{3\pi^3 (e/m)(NI)^2} = 48.5\,\frac{V}{N^2 I^2} \tag{29}$$

$$\theta = \sqrt{2\pi^2 \times 10^{-14}\frac{e}{m}}\,\frac{NI}{\sqrt{V}} = 0.19\,\frac{NI}{\sqrt{V}} \quad \text{radians} \tag{30}$$

To a fair approximation, eq. (29) also applies to a short coil of mean diameter d, of N turns, and carrying current I. Figure 17 is a plot of eq. (29) giving, graphically, the focal length

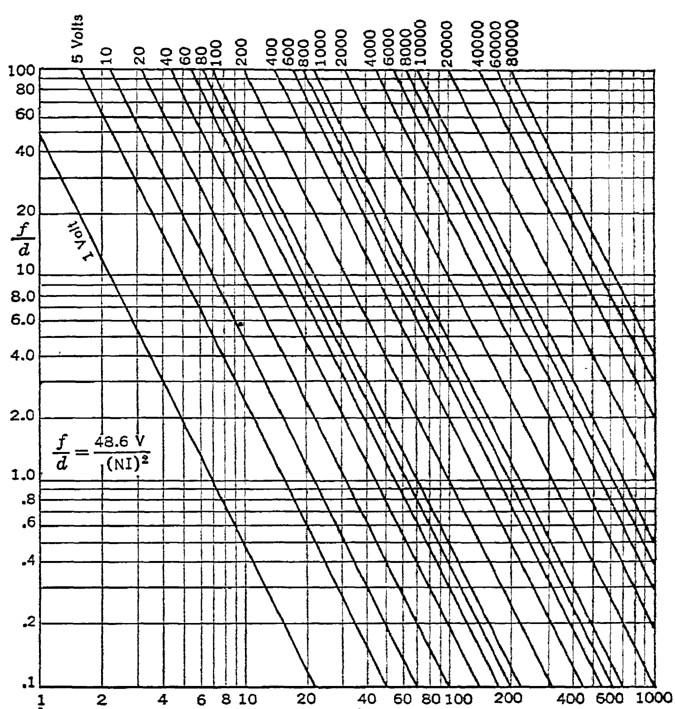


FIG. 17. The Focal Length of a Short Coil as a Function of Ampere-turns and Beam Voltage

as a function of voltage and ampere-turns. Figure 17 may be used, for estimating purposes, in the case of an iron-encased coil if the ampere-turns are small and if d is taken as the clear diameter of the pole pieces. Since the iron-clad coil concentrates the field more than the air coil, the iron coil produces the smaller image rotation.

Besides the usual 5 third-order aberrations of spherical aberration, coma, astigmatism, distortion, and curvature of field, the images produced by magnetostatic lenses usually suffer from three other aberrations. These are generally known as anisotropic distortion, anisotropic curvature of field, and anisotropic coma. Anisotropic distortion is often called the "S effect," since this aberration distorts a straight radial line on the object into an elongated letter S on the image.

16. ELECTRON PRISMS

A uniform electrostatic field and a uniform magnetostatic field constitute the two basic types of electron prisms. Electron prisms are used for deviating (deflecting) beams of electrons of uniform speed as in cathode-ray tubes and pick-up tubes, and for dispersing a beam of charged particles so as to separate particles of differing mass or speed as in the ion trap of a cathode-ray tube and mass spectrograph.

Electrostatic prisms are generally formed by the approximately uniform field between the charged plates of a condenser (see Fig. 18).

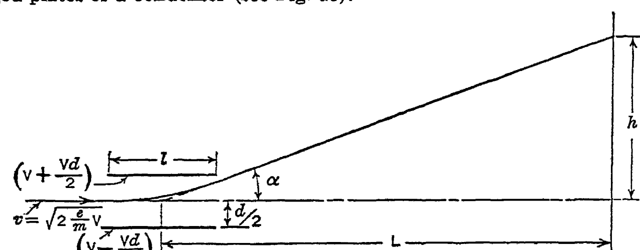


FIG. 18. Deviation (Deflection) of an Electron Beam by an Electrostatic Prism (Deflecting Plates)

A charged particle moving initially perpendicularly to a uniform electrostatic field, E, with a velocity v (corresponding to a potential drop V), follows a parabolic trajectory while in the field. On leaving the field the charged particle is deviated into the direction of the field (see Fig. 18) through the angle α given by

$$\tan \alpha = \frac{lE}{2V} = \frac{lVd}{2dV} \tag{31}$$

where l is the extent of the field or approximately the length of the plates, d the distance between the plates, and Vd is the difference in potential between the plates. The apparent center of deflection is located approximately at $l/2$ from the ends of the plates, and the amount of deflection at the distance L from the center of the plates is

$$h = \frac{lL}{2d}\frac{Vd}{V} \tag{32}$$

Magnetostatic prisms are generally formed by the approximately uniform field between two current-carrying coils or pole pieces of a magnet.

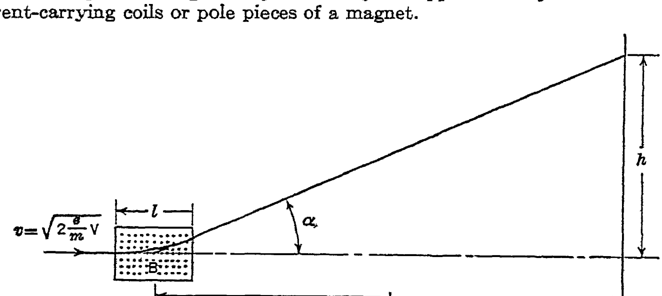


FIG. 19. Deviation (Deflection) of an Electron Beam by a Magnetostatic Prism (Deflecting Coils)

A charged particle moving initially perpendicularly to a uniform magnetostatic field with the velocity v follows a circular trajectory in a plane at right angles to the field. On leaving the field the charged particle is deviated in a direction perpendicular to the field (see Fig. 19) through the angle α given by

$$\sin \alpha = \frac{\sqrt{e/m}\, lB}{\sqrt{2V}} \tag{33}$$

which becomes in the case of electrons

$$\sin \alpha = 2.97 \times 10^5 \frac{lB}{\sqrt{V}} \tag{34}$$

The apparent center of deflection is located approximately at the center of the field, and the amount of deflection at the distance L from the center of the field is

$$h = L \tan \alpha \cong 2.97 \times 10^5 \frac{lLB}{\sqrt{V}} \tag{35}$$

17. GENERAL THEOREMS ON ELECTRON OPTICAL SYSTEMS

Electron optical systems generally consist of a cathode lens and one or more electrostatic and/or magnetostatic lenses and electron prisms (deflecting plates or coils).

The following general theorem applies to an electron optical system using electrostatic and magnetostatic elements. The trajectory of a charged particle remains similar, as long as the quantity $(e/m)(B^2L^2/V)$ is kept unchanged. Thus, if the voltages on all the electrodes are increased by a constant factor n, it is necessary to increase B by the factor $\sqrt{n}$ in order to keep the trajectory the same. If all the linear dimensions (L) of the system, i.e., of all the electrodes, coils, object distance, image distance, etc., are increased by the factor n, it is necessary either to increase all the voltages by n^2 or to decrease B to $1/n$ in order to keep the trajectory similar.

The following theorems apply to any purely electrostatic electron optical system:

1. The trajectory is independent of e/m. Hence all like charged particles emitted by a cathode will be identically focused and deflected by any electrostatic optical system.
2. The trajectory of any charged particle emitted by a cathode is unchanged if the voltages on all the electrodes are increased by the same factor.
3. If all the linear dimensions of all the electrodes are increased by a constant factor n, the trajectory remains unchanged (if measured in units n times larger).

All electron lenses which are free of space charge and conductors within the field of the lens and are bounded by uniform potential regions on both sides are always convergent, i.e., will form a real image of an object located beyond the focal point of the lens.

The maximum current density, J_2, obtainable in an electron spot or image, regardless of the electron optical system employed, is given by

$$J_2 = J_1 \frac{eV_2 + kT}{kT} \sin^2 \alpha_2 \cong J_1 \frac{11{,}600}{T} V_2 \sin^2 \alpha_2 \tag{36}$$

where J_1 is the specific emission of the cathode, kT is the initial kinetic energy of an electron emitted by a thermionic cathode at the absolute temperature T, $k = 1.38 \times 10^{-23}$ erg per degree, $eV_2 + kT$ is the final kinetic energy, and α_2 is the angle of maximum convergence of the electron trajectory at the spot or image.

Equation (36) may be deduced from the fact that, for a perfect optical system which accepts and focuses all the electron current emitted by a thermionic cathode, $J_2 = J_1/m^2$ (m = magnification), and $m = \frac{v_1}{v_2}\frac{\sin \alpha_1}{\sin \alpha_2}$ (see eq. [41], p. 14-08), where v_1 and v_2 are the initial and final velocities of an electron and $\alpha_1 (= 90°)$ is the angle of maximum convergence at the cathode.

REFERENCES

1. Bruche and Scherzer, *Geometrische Elektronenoptik.* Julius Springer, Berlin (1934).
2. Maloff and Epstein, *Electron Optics in Television.* McGraw-Hill (1938).
3. Meyers, L. M., *Electron Optics.* Chapman & Hall, London (1939).
4. Zworykin, V. K., et al., *Electron Optics and the Electron Microscope.* John Wiley (1945).

SECTION 15

ELECTRO-OPTICAL DEVICES

ELECTRO-OPTICAL DEVICES

PHOTORESPONSIVE DEVICES

By Herbert E. Ives

1. RADIATION, PROPERTIES, AND MEANS OF DETECTION

Photoelectric tubes, or cells, belong to the general class of instruments responsive to radiant energy, or *radiation*. An intelligent use of such instruments demands a knowledge of the more important characteristics of radiation, which may be summarized as follows: Radiant energy consists of electromagnetic waves, of a vast range of wavelengths, extending from radio waves, many meters long, to x-rays and γ-rays less than a millionth of a millimeter in length (see Section 11, article 5). The portion of this extended spectrum which is effective in photoelectric tubes and similar devices is confined to a small region centering around that group of wavelengths commonly called light, or the *visible spectrum*, which extends roughly from 0.8 micron (1 micron = 10^{-4} cm) to 0.4 micron. (Other units of measurement frequently encountered are the angstrom unit, 10^{-8} cm, and the millimicron, 10^{-7} cm.) The adjacent region of longer wavelength is called the *infrared;* that of shorter wavelength, the *ultraviolet*. The radiations in the visible spectrum vary greatly in their ability to produce the sensation of light, the relationship between energy value, or radiant intensity, and luminous value being indicated by the *luminosity curve* of the spectrum which has a maximum at approximately 0.5 micron, dropping to zero at the ends of the visible spectrum. Radiant energy is measured in energy units, namely, ergs or joules; and radiation or radiant flux is measured in ergs per second or watts: The *irradiation* of a surface, such as that of a phototube, is measured in ergs per second incident per unit area, or watts per unit area. Luminous flux is radiant flux evaluated according to its luminosity (the integral of the product of the radiation spectrum and the luminosity curve of the spectrum). Its unit is the lumen. One lumen is radiated in unit solid angle by a point source of luminous intensity of 1 candle. The *illumination* of a surface is measured in lumens incident per unit area. The *meter-candle* or *lux* is an illumination of 1 lumen per square meter.

All instruments which respond to radiation do so because they absorb some of the energy of the radiation and transform it into some form susceptible of practical measurement or utilization. *Thermal* instruments become heated by the incident energy, and indicate by changes of dimensions, position, or the production of electrical currents. *Photoelectric* instruments depend primarily upon the fact that the incidence of radiation on matter causes the emission or release of electrons, setting up electromotive forces or causing electrical currents to flow.

CLASSIFICATION OF PHOTORESPONSIVE DEVICES. It is convenient to divide photoresponsive devices into two broad groups, namely, *thermal* devices and *photoelectric* devices. In the group of thermal devices are included thermojunctions, bolometers, and radiometers. In the group of photoelectric devices are included photoemissive cells, photoconductive cells, barrier photocells, and photovoltaic cells.

The *thermal* instruments in general respond to a wide region of the spectrum, corresponding to the thermal absorbing power of the materials of which they are constructed or with which they are coated. Since they are frequently coated with a "black" of high and substantially uniform absorbing power through the infrared, visible, and ultraviolet portion of the spectrum they are often termed "non-selective." They are generally less sensitive than the photoelectric instruments in the spectral region to which the latter respond; they are much slower in response and, being usually of relatively low resistance, are not easily adapted to methods of electrical amplification.

The *photoelectric* instruments are in general sensitive to relatively narrow regions of the spectrum (are strongly selective), but they are far more sensitive in these regions than the thermal instruments. They have also the great practical advantage that they are readily adaptable to the various modern methods of electrical amplification.

2. THERMAL DEVICES

THERMOJUNCTIONS. When two different metals are joined and the junction is maintained at a temperature different from the rest of the circuit, a potential is generated whose magnitude is dependent on the temperature difference and on the materials of the junction. When this difference is very small the potential will be proportional to the temperature. Bismuth and antimony and some of their alloys give large potentials compared with most other metals, and many thermopiles have been made of them. Good, durable combinations of high sensitivity commonly used at present are bismuth against silver, bismuth-tin against bismuth-antimony alloys, and Constantin against Manganin.

Other requirements beyond high thermoelectric power are a low internal resistance and heat capacity and low thermal conductivity between the hot and cold junctions, and in practice a compromise has to be made between these factors. Thermoelements are used in three principal forms: the single junction, the multiple-junction thermopile, and the radiomicrometer, the last being a galvanometer with a thermojunction built directly into, and forming a part of, the moving coil. Multiple junctions are preferable where the illumination extends over some area; for example, where spectral lines are to be measured a linear arrangement of couples is often used, a diagram of which is shown in Fig. 1. The warm junctions, *a*, which receive the radiation to be measured, are commonly flattened or covered with small, thin disks or squares of metal to present a large receiving area. These disks are usually blackened, for instance, with a layer of lampblack, in order to make them uniformly absorptive to a long range of wavelengths. The connection to the cold junctions, *b* and *b'*, must conduct as little heat as possible and at the same time not have too high an electrical resistance. In order to secure the maximum stability it is desirable to shield these cold junctions thoroughly from scattered radiation and to maintain their temperature substantially equal to that of the walls of the surrounding enclosure. This is accomplished by attaching fins of relatively large area to the cold junctions.

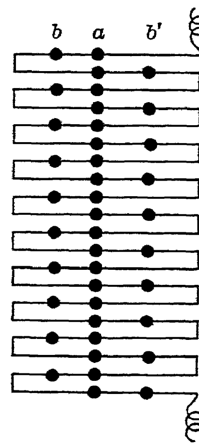


Fig. 1. Linear Thermopile

The galvanometer should have a resistance not far from that of the thermopile, usually from 10 to 50 ohms. There is no advantage in increasing the number of junctions beyond a certain point, and in fact with a given flux the overall sensitivity may be the largest with a few junctions of small heat capacity especially if they are enclosed in a vacuum which eliminates the loss of heat by conduction and convection to the atmosphere and reduces the disturbing effect of drafts. The sensitivity of thermopiles is expressed in volts per unit of irradiation, in some cases, in terms of the response from a stated source at a given distance, commonly a Hefner candle at 1 meter. Values for some of the standard thermopiles are given in Table 1.

Table 1. Sensitivity of Thermopiles

	Ohms Resistance	Volts per Erg per Second per Square Centimeter
Moll linear	20	0.75×10^{-8}
Hilger linear	10	0.7×10^{-8}
Moll large surface	50	5.0×10^{-8}
Moll sensitive vacuum couple	45	0.46×10^{-8}
Moll quick vacuum couple	20	0.13×10^{-8}
Coblentz vacuum couple	14.8	0.7×10^{-8}

For the sensitivity in watts per square centimeter multiply the above figures by 10^7, and for gram calories per second per square centimeter multiply by 4.186×10^7. One Hefner standard lamp with 14 by 50 mm diaphragm opening gives, at 1 meter, 9.6×10^{-5} watt per sq cm or 9.6×10^2 ergs per sec per sq cm.

If a thermopile is to be used with weak radiation, as it frequently is, it is very important that the thermopile and the galvanometer be adapted to each other in order to secure a high sensitivity of the system, i.e., the ratio of the scale reading to the flux. Besides high voltage sensitivity of the pile previously mentioned, other requirements are that the galvanometer should have a low internal resistance, high voltage sensitivity, and a critical damping resistance equal to that of the pile inclusive of the leads. Such galvanometers are made by the Leeds and Northrup Company, Philadelphia, and Kipp and Zonen, Holland.

The sensitivity of the system can sometimes be greatly increased by concentrating the radiation on the junctions and also by using one of the numerous optical devices for amplifying the galvanometer deflections. These are arrangements whereby a light beam reflected from the galvanometer mirror actuates a photocell, the output of which in turn produces an increase in the deflection or operates a second galvanometer. In the Moll thermorelay made by Kipp and Zonen, or the similar Zernike differential couple, the reflected light beam determines the relative temperatures of two opposing thermojunctions, the net output of which controls a second galvanometer. In this manner the readings can be increased up to the inherent instability of a galvanometer, but care must be taken to preserve the linearity of the entire system.

BOLOMETER. The bolometer is essentially a sensitive resistance thermometer of very small heat capacity; that is, the electrical resistance of a fine wire or strip of metal is increased by the heating due to the radiation. It was used extensively before the development of the modern high-sensitivity, quick-acting thermopile, which is more convenient and stable under ordinary conditions. In its practical form two bolometer elements form two arms of a Wheatstone bridge, which the radiation unbalances. It is essential that a material such as platinum be used which has a high temperature coefficient of resistance as well as a low heat capacity and conductivity. Unless unusual precautions are taken the entire bridge network is subject to temperature fluctuations which render the readings uncertain. The sensitivity of bolometers is about one-millionth of a degree per millimeter deflection of a galvanometer used without intermediate amplification. For example, a bolometer of 2.8 ohms resistance and using 40 mils current gave a deflection of 45 cm when exposed to 1 candle at 1 meter with a galvanometer sensitivity of 1.5×10^{-10} amp per mm, the scale being at 1 meter.

THERMISTOR BOLOMETER. A more recently developed form of bolometer is one made of thermistor materials, that is, semiconductors whose resistance varies rapidly with temperature. Combinations of oxides of nickel, manganese, and cobalt change their resistance about 4 per cent per degree centigrade, or about ten times as much as platinum. The oxides are prepared in the form of thin flakes cemented to glass or quartz. A typical flake 3 mm long, 0.2 mm wide, and 0.01 mm thick has a resistance of 4×10^6 ohms and with 250 volts applied gives a sensitivity of 300 volts per incident watt or 18 volts per watt per sq cm. The spectral response is determined by the optical properties of the oxide constituents, which may show regions of relative transparency in the infrared. Because of the high resistance the output of the thermistor bolometer is well adapted to electrical methods of amplification.

RADIOMETER. The Nichols radiometer is a self-contained instrument consisting of two similar vanes of blackened mica or platinum on a horizontal arm and suspended in a vacuum. It is a development of the toy known as Crookes' radiometer frequently seen in optical shops, which consists of four vanes, each blackened on one face, on arms balanced on a needle-point, which rotate when illuminated. The behavior of the radiometer is dependent on the gas pressure; at higher values the blackened sides of the vanes are drawn in turn toward the window; at lower pressures the warming of the blackened faces by the radiation causes the residual gas molecules to rebound from their surface directly to the cooler window and push the vanes away from it and the radiating source. All practical instruments have the gas pressure so adjusted that they work by the latter method. The Nichols radiometer is used by measuring the deflection of the vanes by means of a small mirror attached to the cross-arm supporting them, its rotation being observed by the usual telescope and scale. The sensitivity of the radiometer to radiation is of the same magnitude as that of the bolometer with a sensitive galvanometer.

SPEED OF RESPONSE OF THERMAL DEVICES. The thermal devices are in general slow to respond to variations in signal strength and hence are not well adapted for following rapidly fluctuating radiation, such as radiation modulated at speech frequencies. The time constant, defined as the interval in which the response declines to $1/e$ value, is for the fastest devices of the order or 5 milliseconds.

3. PHOTOEMISSIVE CELLS

STRUCTURE. In photoemissive cells an electropositive metal surface is placed in a highly evacuated enclosure, usually of glass or quartz, together with another metal plate, the electropositive material constituting the sensitive cathode, the other plate the anode, and both plates are connected with terminals led through the glass. The action of the cell is as follows: When light falls on the cathode, electrons are emitted into the space above. These pass over to the anode, and, if the terminals outside the cell are connected through a current-measuring device, a current is observed which varies with the total illumination. Although some current will flow without a battery being connected in series, it is common practice to use one, and then the cells act primarily as valves, the illumination controlling the amount of current which is permitted to pass.

Practically all photoemissive cells use alkali or alkaline-earth metals as their light-sensitive materials, and the structure is largely influenced by the problem of introducing these metals into the glass or quartz enclosure. In early types of photoemissive cells an alkali metal, such as sodium or potassium, was introduced in molten form into a simple spherical bulb provided with one leading-in wire in contact with the pool of alkali metal, and a

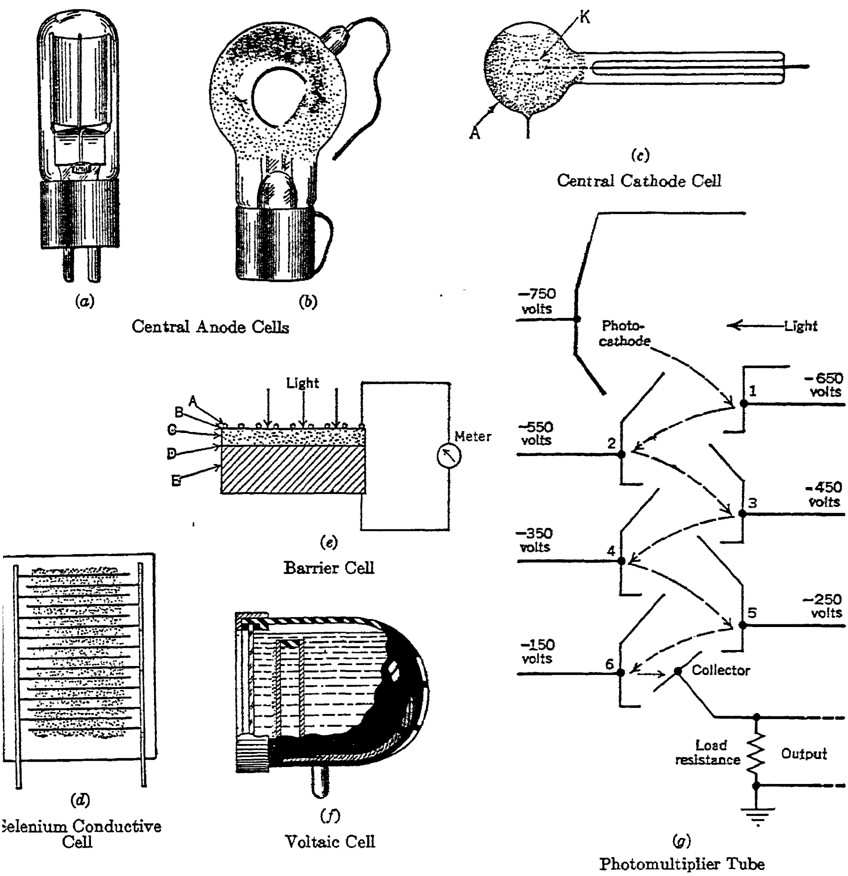


FIG. 2. Types of Photo-responsive Cells

second wire to serve as an anode. This type has been almost entirely superseded by cells in which the alkali metal forms a thin film upon some other metal. Two general types of structure are common, respectively, the central-anode and the central-cathode types. The central-anode type is represented by cells (*a*) and (*b*), Fig. 2. In these the alkali metal has been distilled upon a metal plate or on the walls of a bulb with an opening left for a window for the light. The central anode consists of a wire loop or grid. The central-cathode type, which is illustrated in (*c*), Fig. 2, consists of a metal plate, on which the alkali metal has been deposited, more or less surrounded by a grid or metal covering on the bulb wall serving as the anode.

ELECTRODE MATERIALS. The most commonly used photosensitive materials are the alkali metals, sodium, potassium, rubidium, and cesium, whose intrinsic sensitiveness increases in the order given. The sensitiveness of the pure metals is, however, far below that of the same metals when given various special treatments, such as being exposed to a glow discharge in hydrogen or being put down upon an oxidized base and given special heat treatments. The most generally used cell at present consists of a silver plate which is oxidized and subsequently exposed to cesium vapor and heat treated. A later development has been the combination of antimony with cesium, which produces a high sensitivity localized in the blue region of the spectrum. Other materials, such as barium and cadmium, are occasionally used when sensitiveness to particular regions of the spectrum is required.

VACUUM AND GAS-FILLED CELLS. Photoemissive cells are commonly made up either as high-vacuum cells or as gas-filled cells. In high-vacuum cells, the photoelectric current consists entirely of the electrons emitted from the cathode. In gas-filled cells, an inert gas such as argon is introduced at a pressure of several tenths of a millimeter of mercury, and, when a sufficient voltage is applied, the photoelectrons by collision with the gas produce ionization, whereby the current is amplified several times. This process of gas amplification is limited by the ignition voltage, at which a visible glow discharge takes place, which is self-perpetuating and will injure the surface of the cathode.

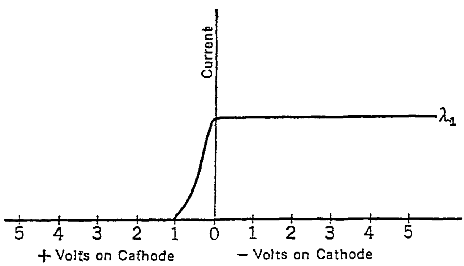


FIG. 3. Current-voltage Relation for a Central Cathode Cell

VOLTAGE-CURRENT CHARACTERISTICS. The relationship between voltage and photoelectric current under constant illumination depends upon the physical structure of the cell, particularly upon the size and arrangement of the electrodes. The structure best adapted for studying the fundamental phenomena of photoelectricity is the central-cathode arrangement. The voltage-current relation for a vacuum central-cathode cell, in which the dimensions of the cathode are negligibly small compared with those of the anode which encloses it as completely as is possible while allowing space for the entrance of light and mechanical parts, is shown in Fig. 3. In this figure, the abscissas represent voltages applied to the cathode and show that, at a certain positive voltage, that is, with a field which opposes the emission of electrons, the photoelectric current makes its appearance. This point, which is a measure of the maximum energy given to the photoelectrons by the incident light, is called the "stopping potential." As the positive voltage is decreased, the photoelectric current increases until it reaches a steady value at the saturation voltage. This voltage will be zero if anode and cathode are of the same material, but it will be displaced by the contact difference of potential between the anode and cathode where the materials are different.

In Fig. 4 is shown the corresponding characteristic for a vacuum central-anode cell. Here, because of the small target presented by the anode, saturation is reached only at high voltages. In Fig. 4 is also shown the voltage-current relationship for a gas-filled cell of the same construction. At low voltages this characteristic is essentially that determined by the structure of the cell, whether it be central anode or central cathode, but at higher voltages the current increases rapidly up to the ignition voltage of the gas which is, in general, of the order of magnitude of several hundred volts. Beyond this point, a sustained discharge occurs with a negative voltage-current characteristic.

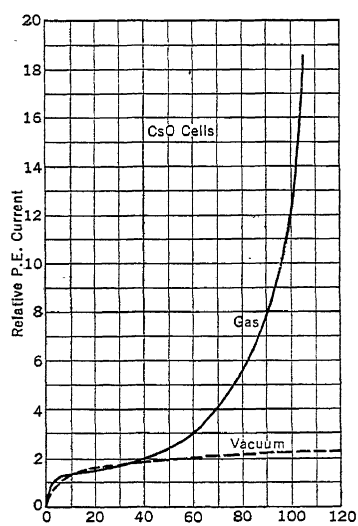


FIG. 4. Current-voltage Relation for Gas and Vacuum Central Anode Cells

ILLUMINATION-CURRENT RELATIONSHIP. With an ideal cell structure, the number of electrons released by the light, and consequently the photoelectric current, are directly proportional to the illumination. In all practical cells, however, this strict relationship is departed from to a slight extent because of charging effects of exposed glass walls and other obscure phenomena. For this reason, photoelectric cells are applicable to precision photometric measurement only if their exact characteristics are determined by experiment, or if a substitution method is used. In gas-filled cells, the illumination-current relationship departs from strict proportionality whenever, as is common, a high series resistance is used either as part of the measuring system or for protection against the occurrence of a glow discharge. The photoelectric current flowing through this series resistance uses up part of the applied potential, thereby lowering the voltage across the cell itself, which accordingly works upon a lower point in the voltage-current characteristic of Fig. 4. Typical illumination-current curves for a gas-filled cell with various series resistances are shown in Fig. 5.

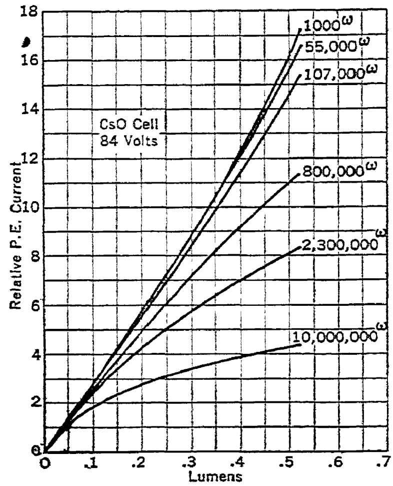


FIG. 5. Current-Illumination Relation for Gas Cell with Various Series Resistances

WAVELENGTH RESPONSE. The photoelectric current per unit of incident radiation varies greatly with wavelength and in different ways, depending on the characteristics of the sensitive material employed.

Figure 6 shows the equi-energy response curves for several typical cells compared with the average eye. These exhibit maxima of emission strongly localized in different parts of the spectrum. The maximum lies in the blue for the potassium hydride cell, in the infrared for the cesium-silver oxide cell, and in the near ultraviolet for the cesium-antimony cell. The type of cell to choose for a given purpose to secure the maximum response depends on the characteristics of the light source used. Sources like the tungsten lamp, whose emission is greatest in the infrared, evoke a maximum response from cells of the cesium oxide type, as illustrated in Fig. 7, where two cells of Fig. 6 are excited by tungsten lamplight instead of an equal-energy spectrum. Daylight and the quartz mercury arc, on the other hand, evoke a greater response from cells whose sensitiveness is farther toward the blue end of the spectrum.

SENSITIVITY. The luminous sensitivity of photoemissive cells is ordinarily defined by their output in microamperes per lumen of steady light. For gas cells which have some

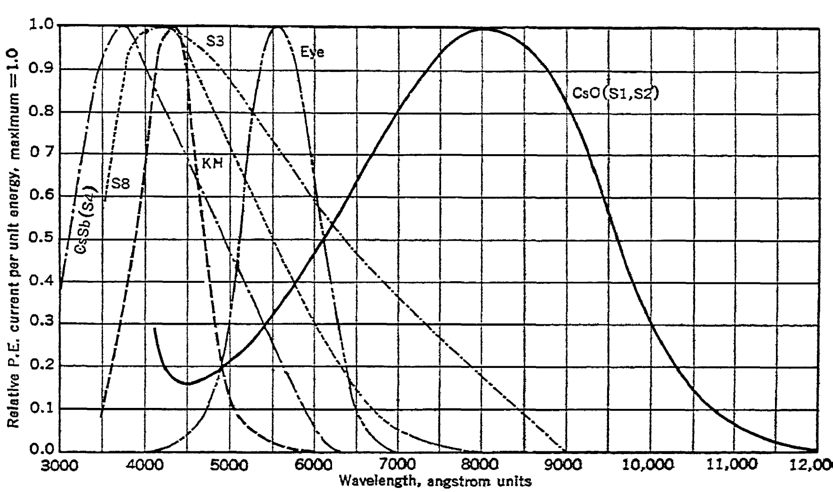


FIG. 6. Equi-energy Spectral Response for Some Types of Blue- and Red-sensitive Cells and the Average Eye

inertia at acoustic frequencies this sensitivity is often stated in terms of the cell's ability to follow a sinusoidally varying light at one or more stated frequencies within this range.

In general the sensitivity depends on the cell voltage, the flux intensity and its color, the resistance in the external circuit, and the distribution of the flux over the cathode if it is non-uniform, as it frequently is to some extent. No allowance is ordinarily made for the glass bulb or any absorbing films on its interior as they must be tolerated in using the cell.

Since the output is dependent on the spectral composition of the light, this should be specified; as tungsten lamps are ordinarily used in rating cells, it is done by stating their color temperature. There is no general agreement on a standard color temperature, but 2870 deg abs has been suggested and is used by some manufacturers. Figure 8 shows how the sensitivities of the red-responsive cesium oxide and the blue-responsive potassium hydride cells vary with this temperature. The sensitivity should not be confused with the total output obtainable from a cell as the filament temperature of the lamp is raised; this output always increases rapidly. The current per lumen of light, however, decreases with highly red-sensitive cells because there is relatively less red to blue or total light. Typical sensitivities are given in Table 2.

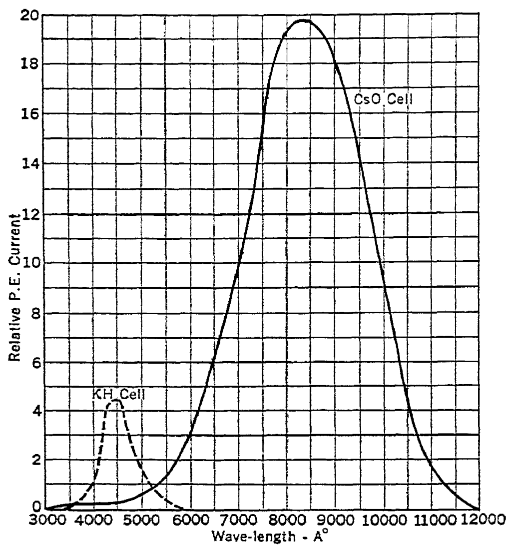


FIG. 7. Tungsten Spectral Response for KH and CsO Cells

The sensitivity of photocells is also defined by the magnitude of their response per unit radiant flux at a given wavelength. For example, the cells in Fig. 6 may be so evaluated by their microampere output per microwatt of irradiation at the wavelengths of maximum response. Typical values for unamplified commercial cathodes are: S1, 0.002 at 7500 Å; S2, 0.002 at 8000 Å; S3, 0.002 at 4400 Å; S4, 0.04 at 3750 Å.

In evaluating photoemissive cells the permanence and stability of the sensitivity must be considered along with its absolute value. Commercial emissive cells are usually permanent and stable provided that excessive voltages are not used (particularly on gas cells)

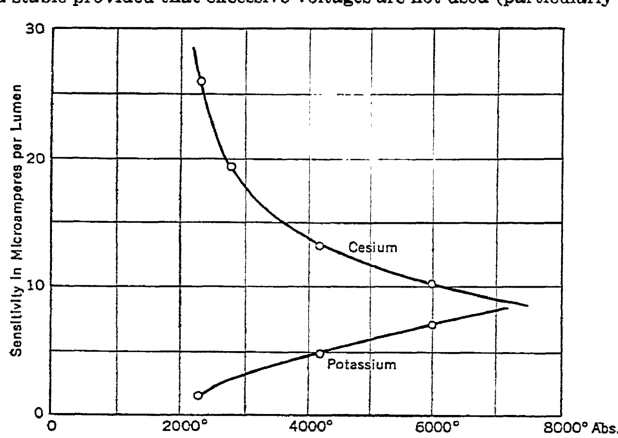


FIG. 8. Variation of Sensitivity with Color Temperature

and provided that they are not subjected to excessive heating or too concentrated illumination. Any ionization even of residual vapor or gas must be avoided for maximum stability so that in precision use anode voltages as low as 20 volts are recommended.

Leakage, due to a conducting film around the stem or to thermionic emission from the cathode, may render high sensitivity more or less useless. Cathodes deficient in red sensitivity, such as cesium-antimony, have greatly reduced thermionic emission.

FREQUENCY RESPONSE. The emission of electrons in the photoelectric effect is practically instantaneous, and accordingly the emission should be capable of following intermittent illumination up to exceedingly high frequencies, such for instance as those required for television. In practice, the frequency response drops off at high frequencies, owing either to the presence of the gas in gas-filled cells or in vacuum cells to the presence of the high series resistance, which is ordinarily used for coupling purposes, or to the electrostatic capacitance of the cell considered as a condenser. Typical frequency-response curves are given in Fig. 9.

Table 2. Sensitivity of Photoemissive Materials

(Tungsten lamp source)

	Microamperes per Lumen *	Filament Color Temperature, deg Kelvin
Potassium hydride	0.5–1.0	2848
Potassium-sulfur	1.8	2848
Sodium-sulfur	2.4	2848
Sodium-sulfur-oxygen	6–8	2848
Cesium antimony (S4)	30–45	2870
Cesium oxide on silver (S1, S2)	10–50	2870
S3	6.5	2870
S5	15.0	2870
S8	3.0	2870

* Intrinsic sensitivity of the cathode; amplification by gas or by secondary emission may increase these figures by several times or by several orders respectively.

In a gas cell the frequency response is dependent on the applied voltage, and if this is near the breakdown the loss of response may considerably exceed the values shown. It is not advisable to exceed the voltage recommended by the manufacturer.

MEASURING CIRCUITS FOR USE WITH PHOTOEMISSIVE CELLS. There are two general methods of measuring photoelectric-cell output: first, the measurement of the current directly; and second, the measurement of the voltage drop across a series resistance. The photoelectric current is measured by inserting a sensitive galvanometer in series with the cell and a battery, and the method is limited in sensitiveness only by the sensitivity of available galvanometers. The voltage drop across a high resistance is measured by means of an electrometer or of special vacuum tubes designed to function in the same manner. For extremely minute illuminations, the high resistance may be made infinite, and the current may be ascertained by the rate at which the electrometer or equivalent device charges up.

A number of d-c galvanometers are made which are suitable for use with photoemissive cells, ranging in sensitivity down from about 10^{-10} amp; this can be extended to about 10^{-12} amp with a device like the Moll thermorelay. The resistance of the galvanometer is immaterial, provided that it has a high current sensitivity, because the resistance of the cell will ordinarily be enormously greater. It is not desirable for several reasons to use an instrument that is much more sensitive than the work demands, but the testing of various types of cells of widely varying characteristics with different colors and intensities of illumination may require a very flexible galvanometer system. In this event, shunts can be used if they are so arranged as not to interfere with the critical damping. A good method is to insert various sections of the damping resistance in the cell circuit. The Leeds and Northrup type 2285-F galvanometer is suitable for this purpose. Where considerable current is available and portability is desired, microammeters of the Rawson or Weston type are convenient, especially when provided with a dial or other means of securing various scale sensitivities.

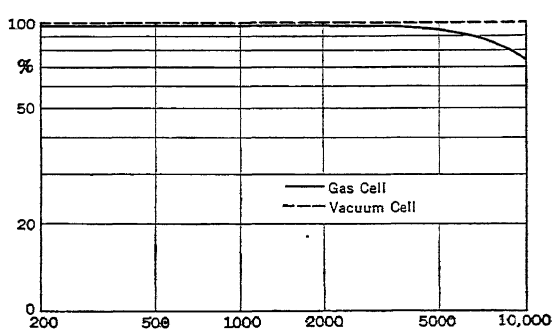


FIG. 9. Frequency Response for Gas and Vacuum Cells

In all cases, great care must be taken to protect any instrument from a breakdown of the cell by inserting sufficient series resistance, at the same time, if possible, retaining linearity of response. For work of the highest sensitivity it is necessary to resort to an electrometer such as the Compton type, the applied potential being secured by the drop across several megohms in series with the cell, or by using the cell as a constant current source and measuring the rate of charging of the electrometer placed in series. Effective use of the electrometer requires a permanent laboratory installation. Where greater portability is desired practically equivalent results can be secured by amplifying the very small currents by an "FP.54 pliotron" and measurement on a moving-coil galvanometer such as the Leeds and Northrup type R. For the methods of using an electrometer see Hughes and DuBridge's *Photoelectric Phenomena*, 1932, pp. 435–444, and for the corresponding technique of amplification see DuBridge, *Physical Review*, Vol. 37, Feb. 15, 1931, pp. 392–400. For descriptions of the more common electrometers see the catalog of the Cambridge Instrument Company, and for galvanometers consult catalogs of Leeds and Northrup Company and Kipp and Zonen.

AMPLIFICATION OF PHOTOEMISSIVE-CELL OUTPUT. Photomultiplier Tube. The initial current produced by illumination of a photocathode may be greatly amplified by utilizing the phenomenon of secondary emission. In the photomultiplier tube the electrons emitted from the cathode are directed by a suitable high-voltage field onto another, usually similar, electrode where they cause the emission of secondary electrons which are greater in number than the impinging electrons. This process may be repeated a number of times. Present commercial multiplier tubes have as many as nine such stages, and even larger numbers have been used in special tubes. The overall current amplification thus produced is of the order of several hundred thousand times. A typical electron multiplier structure is shown in (*g*), Fig. 2, and commercially available models are listed in Table 3, p. 15-16.

Circuits for Amplifying Photoemissive-cell Output. The photoemissive cell, because of its exceedingly high internal resistance, is admirably adapted for use in connection with vacuum-tube amplifying devices. In the simplest arrangement, the electrometer described in the previous section is replaced by the grid of a three-electrode tube, and the potential acquired by it modulates the current through the vacuum tube, which may be further amplified by successive stages.

Circuits for amplifying photoemissive-cell output are determined by the type of application of the cell. These can be put largely into three classes: trigger operation, d-c linear operation, and a-c linear operation. The first class is concerned with merely a qualitative response, and there is no particular requirement for linearity. A large number of uses come under this classification, such as the operation of relays in various counting and sorting processes. The usual circuit for amplification consists in feeding the voltage drop across the cell load into a thyratron tube which in turn actuates a relay (see Section 21). The necessity for supplementary amplification will depend on the light variation available and the marginal requirements of the relay. It is necessary to provide for the release of the thyratron, and this can be easily done by using alternating current on its cathode or interrupting the direct current. The sensitivity of trigger systems can ordinarily be increased by the use of large load resistances in the cell circuit.

A method of use closely allied to the preceding in the characteristics demanded of the cell is as a null device in substitution photometry, where the only requirement is a suitable recorder with enough amplification to give the necessary precision. If the test and comparison lights are rapidly alternated on the cell, the method permits of a-c amplification with its attendant advantages of great efficiency and simplicity, the match being given by zero a-c output.

The second class, d-c linear operation, covers the direct-reading method of photometry, an example of which is the recording of daylight intensity. This method has the disadvantage of requiring stable cells of reproducible, linear characteristics, a requirement not always easy to meet for precision photometry. For many purposes, however, the requirements are not rigid and commercial cells are suitable for the purpose. Several methods of amplification are possible. One is to use straight d-c resistance-coupled amplification, preferably at low cell currents, with an electrometer tube as the first stage. In order to minimize the tendency to instability inherent in d-c amplification, use is sometimes made of balanced d-c amplification. At low light intensities these direct methods require much care to guard against leakages in and around the cell, and, for some purposes, specially constructed cells are necessary. Another precaution that must be taken is to insure the linearity of response of the amplifier. A third method of amplification that is sometimes applicable is to interrupt the illumination of the cell and use a-c amplification (see Section 7). This has the advantages of greater stability, of minimizing leakages, and of securing the efficiency of interstage coupling by transformers.

Such uses of photocells as for picture transmission and sound pictures may be classed as a-c linear operation. Here the response must be not only linear but also uniform over a range of frequencies which may, as in television, be very large. This imposes certain restrictions on the load impedance in the cell circuit. Since a large value is necessary in order to secure a high voltage output into the amplifier, the shunting capacitance becomes very important and severely limits the useful value of the impedance that can be used.

Noise Limit to the Use of Amplification. Broadly speaking, any photoelectric current, however minute, can by successive amplification be raised to any desired high value. A limit is set to the effectiveness of this process by the noise in the cell or its associated circuits, which is amplified along with the signal. The significant specification of sensitivity of a photocell thus becomes the signal which can override the noise. Noise in a photocell exists because of natural fluctuations of current at low values and by the thermal agitation of electrons in the coupled resistances, and it is a function of temperature, frequency, and band width. In addition there are many extraneous sources of noise from such causes as the ionization in gas tubes, interference, microphonic contacts, power supplies, and dielectric leakages. These can usually be diminished to secondary importance by careful shielding and design of the circuits. Where considerable amplification of weak photocurrents of wide bandwidth is required there will be a gain in the signal-to-noise ratio by using a multiplier phototube for the initial stages.

4. PHOTOCONDUCTIVE CELLS

The fact that light could directly change the electrical resistance of a substance was first discovered about 1880 by observation of the effect in metallic selenium, and since that time some 2000 papers have been published concerning its behavior and use. Nevertheless, the mechanism whereby light releases electrons remains obscure. Furthermore, many of its characteristics depend to a considerable extent on the method of construction as well as the conditions under which they are measured. Two types of construction have been used for such cells: one in which a thin layer of selenium is sandwiched between two electrodes, one of which must be translucent to permit illumination of the layer; and one in which two interlocking metallic grids or combs are bridged by a layer of selenium. Modern cells are usually of this second type, although the barrier cells described below use the first. An example of a conductive selenium cell of comb construction is shown in Fig. 2(*d*). Other materials which exhibit similar properties are thallium sulfide and lead sulfide. The former is used in a cell known commercially under the name of "Thalofide." Both these newer cells possess properties which render them superior to selenium.

Current-illumination Relationship. If a selenium cell is placed in series with a battery and meter and is illuminated with increasing intensity, the resulting current will usually be of the shape shown in Fig. 10. In general the change in conductance, G, follows the equation

$$G = \frac{1}{r_i} - \frac{1}{r_0} = gF^x$$

where g and x are constants, F the light flux, r_0 the dark resistance, and r_i the light resistance. The constant x is frequently about 0.5, so that the photocurrent varies approximately as the square root of the illumination. The dark resistances of different grid cells vary greatly but usually are from 100,000 up to several megohms; those of the sandwich type are much lower and may be only a few hundred or thousand ohms. The current-illumination relation of thallium sulfide differs from selenium in not being curved as strongly toward the illumination axis.

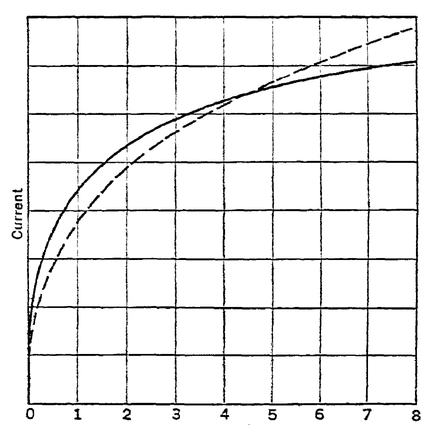


FIG. 10. Current-illumination Response for Selenium Conductive Cell

Photoconductive cells are made to operate on a variety of applied voltages, and in general the recommendations of the maker should be followed. If not hermetically sealed the cells should be protected from excessive moisture and corrosive gases such as sulfur fumes. High illumination also causes deterioration of some cells.

WAVELENGTH RESPONSE. Typical curves of the spectral response for both selenium and thallium sulfide are shown in Fig. 11. It is characteristic of the former to have a peak at 7000 or 7500 angstroms, the sensitivity up through the visible region being variable with different cells and frequently rising again as the ultraviolet is approached. Thallium sulfide is much more infrared-sensitive than selenium, with a maximum around 10,000 angstroms, after which it falls off rapidly, whereas lead sulfide has a maximum at 25,000 angstroms and falls to 20 per cent at 4000 and 33,000 angstroms.

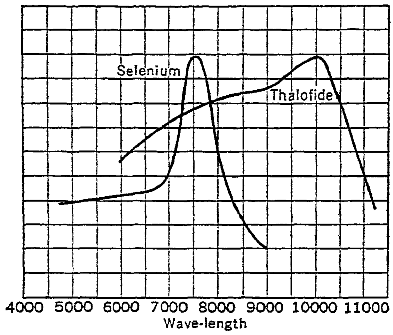


FIG. 11. Spectral Response for Selenium and Thallium Sulfide Conductive Cells

FREQUENCY RESPONSE. Many oscillographic observations have been made on the speed with which the photocurrent builds up when a selenium cell is suddenly illuminated and on the rate of decay of the current when the light is removed. This method of observing its inertia, however, is ordinarily not so useful as the method of measuring its response to continuously interrupted light of known frequencies. Figure 12 shows such a measurement starting at very low and extending to nearly 10,000 interruptions per second, the ordinates being the a-c response relative to the flat portion as unity. The most recently developed Thalofide cells fall off much more slowly with frequency and are useful up to the lower voice frequencies; the frequency performance of lead sulfide is notably better than that of thallium sulfide, remaining practically constant up to 5000 cps.

SENSITIVITY. Photoconductive cells are commonly rated by their ratio of dark-to-light resistance at a stated illumination and voltage. Thus a cell may be stated to have a ratio of 6 at 100 ft-c using a battery of 100 volts. As there is no agreement on a standard light intensity or color, care must be exercised in comparing cells from different manufacturers. The d-c sensitivities may be divided into two classes according to whether the load in series with the cell is a relay to be operated directly, or a resistance, the voltage drop across which is in turn required to operate the grid of a thermionic tube. In the latter case the expression for the voltage sensitivity, σ, of selenium is

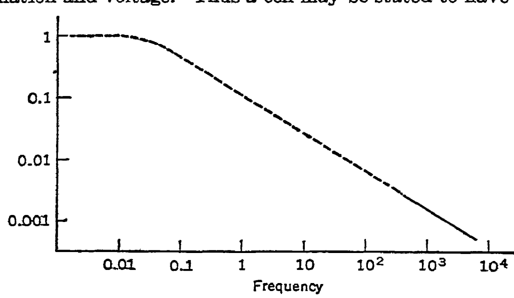


FIG. 12. Frequency Loss for Selenium Conductive Cell

$$\sigma = \frac{dE_g}{dF} = \frac{Er_ig}{8\sqrt{F}} = \frac{Er_0g}{8\sqrt{F}(1 + r_0g\sqrt{F})}$$

where E_g = voltage across load effective on grid.
E = battery voltage.
F = light flux.
r_0 = dark resistance of cell.
g = constant of the cell, its light conductance being $g\sqrt{F}$.
r_i = light resistance of cell.

This is the equation for the maximum voltage sensitivity at light flux F, where the load resistance is equal to the light resistance of the cell r_i, under the condition of operation. If the load is a relay to be operated directly in series with the cell the ampere-turn sensitivity must be used instead of the above. This is equal to the voltage sensitivity given above multiplied by $1/t$, where t is the resistance per turn of the relay.

If a photoresistance cell is operated between dark and a given light intensity as the limits, the maximum voltage change across the load resistance, r_a, is secured when

$$r_a = \sqrt{r_0r_i}$$

the voltage change being

$$Er_a\left(\frac{1}{r_i + r_a} - \frac{1}{r_0 + r_a}\right)$$

The same condition for r_a applies for the current-turn sensitivity, and its change is equal to the voltage change multiplied by $1/t$.

AMPLIFICATION OF PHOTOCONDUCTIVE-CELL OUTPUT. The amplification of the output of these cells is fundamentally the same as for the emissive cells, and the same types of amplifiers can be used. Since conductive cells have much lower impedances, it is possible in many cases to match the load to the cell and thereby secure the maximum efficiency of operation. Since the cell resistance may vary rapidly with illumination, care must be taken that the match be made at the average light intensity at which the cell is to be operated. If an intermittent or variable illumination is used such that the cell must respond to some range of frequencies it may be necessary to equalize the output in order to preserve fidelity of reproduction. Since there is an increasing loss as the frequency of response is increased, it is necessary to compensate for it by introducing attenuation of the lower frequencies by a filter network at some convenient point in the amplifier. The highest frequency at which one desires to work will then determine the effective loss of a cell.

5. BARRIER PHOTOCELLS

STRUCTURE. The fact that, under certain conditions, photoconductive selenium cells could produce an emf on illumination without any applied potential has been known since the late nineteenth century. The discovery of the effect in cuprous oxide revived interest in it and led to the development of cells of commercial importance. They are of the sandwich type of construction referred to above and illustrated in Fig. 2(*e*). A photosensitive material such as selenium or cuprous oxide *C* is formed on a suitable metallic base *E* and covered with a translucent conductor such as a wire mesh *A* or a thin metallic film. If the photosensitive layer is itself partially transparent, as, for example, cuprous oxide, a photo emf may appear at junction *D*, where it is called a "back-wall effect," or at *B*, where it is called a "front-wall effect." The location and degree of sensitiveness are dependent on the method of preparation, the heat treatment, rate of cooling, gas content, and treatment of the boundary surface. If an attempt is made to pass current across such a photosensitive boundary by inserting a battery of a few volts in the meter circuit, it is customarily found that the current can flow much more easily in one direction than in the other, and this directional resistance behavior or rectification is illustrated in Fig. 13 for a typical commercial cell in the dark, the unit here being of the front-wall type. As a photocell without the external battery, the effect of illumination is to make the top become negative and the base positive; that is, the electrons released by the light flow in the high-resistance direction. If the active layer is at the back wall *D*, Fig. 2(*e*), the behavior is primarily the same except that the polarities are reversed with respect to the top and bottom. In this case the light must penetrate much more material, which reduces the optical efficiency and alters the spectral response curve.

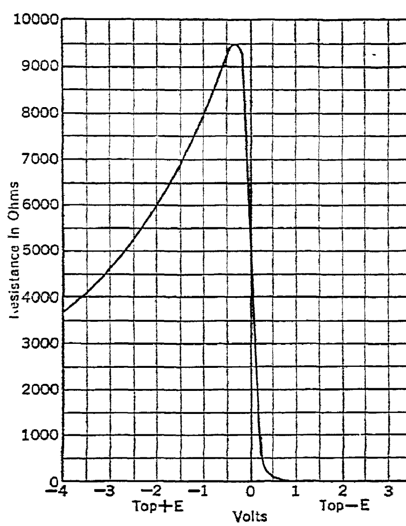


FIG. 13. Dark Resistance-voltage Relation for Selenium Barrier Cell

When a barrier cell is illuminated, its dark resistance at zero applied voltage is decreased by the photoconductive effect in the layer, and at high intensities it may be much reduced for a cell of the characteristics shown in Fig. 13.

ILLUMINATION RESPONSE. The current-illumination response of a typical cell is shown in Fig. 14 with various series resistances. With very low resistance or short-circuited current the relation is linear or very nearly so, gradually becoming more curved as the resistance is increased until the open-circuit voltage relation is reached. Figure 15 shows the two extremes for comparison. Care must therefore be taken not to use too much series resistance if a linear response is desired.

WAVELENGTH RESPONSE. In Fig. 16, *A* shows the spectral response for a typical selenium cell; *B* and *C* are for front- and back-wall cuprous oxide cells, respectively. In general, cells of this type have most of their sensitivity in the visible region. Back-wall cells of cuprous oxide, however, are deficient in this region owing to the absorption of the red oxide so that their response is confined largely to the visible red and some distance beyond into the infrared. Cells equipped with optical filters to make their response closely that of the eye are now supplied by manufacturers.

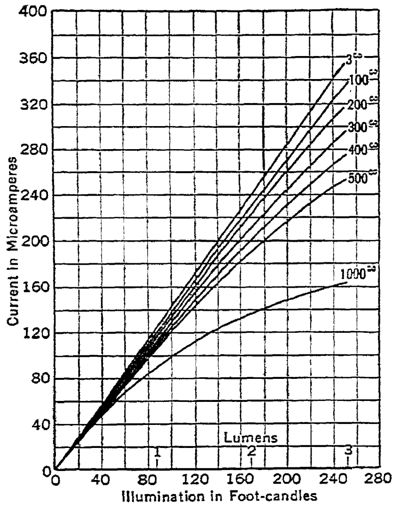


FIG. 14. Current-illumination Relation of Barrier Cell with Different Series Resistances

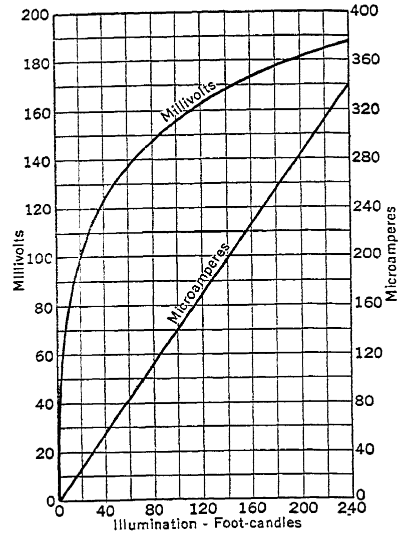


FIG. 15. Short-circuit Current and Open-circuit EMF-illumination Relations for Barrier Cell

SENSITIVITY. Barrier cells are rated according to their microampere-per-lumen output, and the same qualifications apply to them as were stated for emissive cells concerning the color of the light source and resistance in series. On account of the warping of the linearity of the current curve by comparatively small resistances, care must be taken in its measurement to use a sufficiently low-resistance meter. It is also frequently useful in certain applications to have a statement of the open-circuit voltage in millivolts per lumen. From Fig. 15 it is clear that this ratio is high at low illuminations and rapidly diminishes as the illumination is raised, and for this reason it is necessary to state the illumination at which the measurement is made.

Calculations of the maximum power and voltage sensitivities are complicated by the fact that the internal resistance decreases with illumination, especially with the selenium type. For this reason, it is necessary to know the characteristics of the individual cell, and it is advisable to consult the manufacturer regarding the particular use to which it is to be put or to determine the proper load experimentally. For the cuprous oxide type the internal resistance is commonly assumed constant, and in this case the maximum voltage sensitivity, dE_g/dF, is equal to Sr, S being the sensitivity constant and r the internal resistance, the load, r_L, being relatively very large. In this case the maximum power sensitivity is $S^2rF/2$, where $r_L = r$. These values of r_L will ordinarily be sufficient to cause departure from linearity of current response so that a compromise must be made by sufficient reduction in the load resistance. Barrier cells are liable to deteriorate, and direct-reading instruments utilizing them such as foot-candle and photographic exposure meters should be checked occasionally and the cell replaced if necessary. Poor contact with the clamping rings may also develop. This

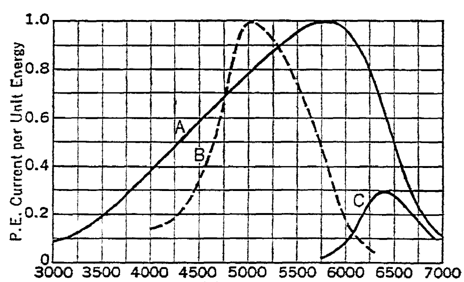


FIG. 16. Spectral Response of Barrier Cells. *A*, selenium. *B*, cuprous oxide, front wall. *C*, cuprous oxide, back wall.

type of cell is subject to a fatigue which causes a decrease in the response when first illuminated. Recovery takes place in the dark, but the reproducibility of readings is somewhat dependent on the duration and intensity of illumination. This type has no dark current.

FREQUENCY RESPONSE. Exact information on the various commercial cells is not available. Barrier cells have a relatively large internal capacitance which diminishes their output with increasing frequency. The "Photronic cell" is stated to give a satisfactory response to light interrupted at 60 cycles per second, and if this output is assumed 100 per cent then at 120 cycles it will be about 58 per cent, at 240 cycles 30 per cent, and at 1000 cycles 6.4 per cent. It is also stated that, if an equalized response is produced up to 5000 cycles, the power level is reduced approximately 35 db. The response of the cell to single interruption is more rapid than that of a relay, so that for this use they may be considered instantaneous.

AMPLIFYING CIRCUITS. If constant illumination is to be used, a d-c amplifier is demanded, which may be troublesome to build and operate. In general it is therefore recommended that the light be interrupted at a low frequency of, for example, about 60 cycles, and that a good a-f amplifier capable of transmitting this frequency efficiently be used.

The remarks previously made concerning the amplification of emissive and conductive cells also apply to this type. Barrier cells have lower impedance than either of these, and are well within the range of practical transformers, so that they can be directly coupled and the output used in an a-c amplifier. Equalization of frequency response may be necessary as with the conductive cells.

6. PHOTOVOLTAIC CELLS

In the middle of the nineteenth century it was discovered that if two similar electrodes of certain materials, such as platinum or silver coated with silver halide, were immersed in dilute electrolytes and one electrode was illuminated, a voltage appeared between them. These are referred to as photovoltaic cells, although some writers broaden this name to include also the barrier cells previously described and refer to them as wet and dry cells respectively. Various combinations of electrodes, coatings, and intervening liquids have been employed, but, apart from experimental studies of the effect, the usual materials are oxides, sulfides, or halides in an acid or inorganic salt electrolyte. Cuprous oxide gives a large effect and has frequently been used in attempts to commercialize this type, the non-sensitive electrode being some durable material such as lead with an electrolyte of lead nitrate in water; see Fig. 2(*f*).

CHARACTERISTICS. The behavior of the open-circuit voltage and the short-circuit current responses with illumination for a cell of the Cu, Cu_2O, $Pb(NO_3)_2$, Pb construction are similar in shape to the corresponding characteristics for barrier cells (see Fig. 15), the former showing a tendency to voltage saturation and the latter being linear. The effect of increasing external resistance is also qualitatively similar; see Fig. 14.

The spectral response of the cuprous oxide cell is high in the visible spectrum, being a maximum in the blue or blue-green.

These cells are subject to polarizing effects which make them more unstable than the other types, and there is a gradual deterioration of the sensitive layer which greatly shortens their useful life. It is claimed that this deterioration can be inhibited to some extent by using a depolarizer such as hydrogen peroxide in the electrolyte to oxidize the free hydrogen which reduces the cuprous oxide.

The response to illumination and the recovery afterward are not as rapid as the cuprous oxide barrier effect.

7. CHOICE OF CELLS FOR VARIOUS PURPOSES

Although the applications of photocells are frequently classified according to marginal or linear operation, this does not have so much to do with the choice of the cell as the method of its use and of the amplification of its output. Ordinarily the selection of a cell is dominated by considerations of high sensitivity to tungsten or daylight, the requirements of precision photometry, a particular spectral response, or convenience, all of which involve the relative evaluation for the purpose at hand of such factors as magnitude of output, frequency loss, fidelity of color and intensity response, permanence, stability, leakage, and absence of external battery.

Table 3. Photoelectric Cells

Type Number	Description	Spectral Response	Typical Sensitivity, microamperes per lumen	Maximum Operating Volts	Cathode Window Dimensions, in.	Maximum I, microamperes	Base	Manufacturer
PJ-22	Vacuum	S-1	20	500	$^{11}/_{16} \times 1\,^{5}/_{8}$	152 per sq in.	M8-074	G.E. Co.
FJ-405 ¶	Vacuum	S-6 Ultraviolet	12	200	1 diam.	62 per sq in.	4102	G.E. Co.
GL-922	Vacuum	S-1	20	500	$^{5}/_{8} \times ^{5}/_{8}$	152 per sq in.	Cartridge	G.E. Co.
GL-929	Vacuum	S-4	45	250	$^{11}/_{16} \times ^{7}/_{8}$	102 per sq in.	M8-046	G.E. Co.
CE *	Vacuum	S-1	A/B, C, D	500	Various			Continental
CE-29	Vacuum	S-4	45	250	0.62 sq in.		Interm. Octal 5-pin	Continental
917	Vacuum	S-2 (S-1) †	20	500 ‡	1 sq in.	30	Tapered Small 4-pin	RCA
919	Vacuum	S-2 (S-1)	20	500 §	1 sq in.	30	Tapered Small 4-pin	RCA
926	Vacuum	S-3	6.5	500	0.4 sq in.	20	Cartridge	RCA
929	Vacuum	S-4	45	250	0.6 sq in.	20	Interm. Octal 5-pin	RCA
R59AV ‖	Vacuum	S-1	40	500	1.37 × 0.62		Tapered Small 4-pin	Rauland
R59BV	Vacuum	S-4	45	500	1.37 × 0.62		Tapered Small 4-pin	Rauland
SR53 ¶	Vacuum	S-1	25	500	1.38 × 0.81	20	410	Westinghouse
NU-R1003	Gas	S-3	25	90	0.2 sq in.	30	Cartridge	National Union
CE *	Gas	S-1	A/B, C, D	90–110				Continental
868	Gas	S-1	65	90	1 sq in.	20	Tapered Small 4-pin	RCA
918	Gas	S-2 (S-1)	150	90	1 sq in.	20	Tapered Small 4-pin	RCA
920	Gas	S-1 (Twin cathode)	75 each	90	0.3 sq in. each	10 each	Small 4-pin	RCA
928	Gas	S-1 (Non-directional)	65	90	0.7 sq in.	15	Small 4-pin	RCA
R59A	Gas	S-1	225	90	1.37 × 0.62		Tapered Small 4-pin	Rauland

R59B	Gas	S-4	250	90	1.37 × 0.62		Tapered Small 4-pin	Rauland
SK63	Gas	S-1	125	90	1.38 × 0.81	20	410	Westinghouse
WL735	Gas	S-1	50	90	1.38 × 0.81	20	4101	Westinghouse
GL-918	Gas	S-1	150	100	11/16 × 1 5/8	102 per sq in.	M8-074	G.E. Co.
GL-927	Gas	S-1	125	90	7/16 × 7/8	101 per sq in.	3313	G.E. Co.
931A	Multiplier	S-4	10^7	1250	15/16 × 5/16	1000	Small Shell	RCA
1P21	Multiplier	S-4	6×10^6	1250	0.25 sq in.	1000	Submagnal	RCA
1P22	Multiplier	S-8	6×10^5	1250	15/16 × 5/16	1000	11-pin	RCA
1P28 ¶	Multiplier	S-5 Ultraviolet	3×10^6	1250	0.25 sq in.	2500		RCA
Photronic No. 1	Barrier	Visible	136	....	1.70 sq in.	....	Prong	Weston
Photronic No. 3	Barrier	Visible	190–405	....	1.70 sq in.	....	Threaded	Weston
Emby	Barrier	Visible	Up to 600	....	Various	....	....	Selenium Corp.
88X565	Barrier	Visible	425	....	1.1 sq in.	....	2-pin for radio sockets	G.E. Co.
97X637	Barrier	Visible	Various	....	1.1 sq in.	....	Unmounted	G.E. Co.

This table is merely to give a few representative samples; catalogs should be consulted for detailed information. For a more complete list of suppliers see *Electronics Buyers' Guide*.

* A series of tubes under the trade name Cetron. The major difference is in the physical dimensions. Three selections of average sensitivity are available: A/B = 35, C = 25, D = 16 for vacuum, and A/B = 300, C = 160, D = 100 for gas.

† S-1 is being generally used for the deep red and near ultraviolet response of C-O-Ag tubes. In some cases S-2 is used for types rated as longer infrared response; if important, tubes should be individually measured regardless of the code.

‡ Anode cap.

§ Cathode cap.

‖ R59 series also available in ultraviolet-transmitting envelope. This and other series by Rauland under the trade name Visitron.

¶ Some ultraviolet tubes not listed are WL767, WL773, WL775, WL789 by Westinghouse and 935 by RCA.

The cesium oxide cell is particularly efficient for use with tungsten light because its maximum sensitivity is near the optimum energy emission of the high-brilliancy gas lamps. This renders it suitable for sound pictures, telephotography, television, and numerous marginal applications such as counting and control operations. The color response is also sufficiently extended over the spectrum to permit its use in many sorting operations and in colorimeters. In some technical applications it is necessary or, at least, highly desirable that the cells be used with infrared light to avoid detection of the beam by the wary or curious, and the cell has sufficient infrared response to permit the tungsten light to be concealed by filters which transmit only this region with very little visibility of the interrupted beam. The cesium-antimony cathode is finding increasing application. Its high efficiency to tungsten light makes it competitive with cesium oxide and considerably superior with bluer modern illuminants. For colorimetric applications, a surface such as the S3 is more suitable because of its broad spectral response through the visible region which more nearly simulates the eye.

The requirements of precision photometry vary greatly according to the intensity and color of the light to be measured and according to whether the measurements are relative or in visual units. For example, in stellar photometry where very little light is available and high amplification is necessary, leakage in and around the cell must be reduced to the minimum and high sensitivity and color response may not be as important. On the other hand, in the ordinary routine photometry of tungsten light, leakage may be relatively unimportant. In precision photometry it is essential that cells be used in a manner which gives an assured calibration for each reading and that their constancy be not assumed without adequate proof. Cells to be used for measurements in visibility units require either a filter to modify their color response or a calibration by a source whose spectral emission is the same as that to be measured. If sources such as tungsten lamps through a moderate range of color temperature are to be measured, requirements of commercial photometry may be met by calibrating with a similar source within the range, combined if necessary with a visual filter that approximately matches the eye.

The increasing use of ultraviolet for therapeutic and photochemical purposes has created a demand for cells to measure intensities in this region, the response for therapeutic purposes being of such shape as to evaluate the radiation directly in erythema units or in time of exposure. This region is from 2800 to 3200 angstroms approximately, the shorter wavelengths of questionable value being excluded by filtering the source. A number of materials are intrinsically sensitive to this region, but it is undesirable to use those of considerable visible sensitivity because of the difficulty of suppressing sufficiently the larger amount of energy in the longer wavelengths even in the mercury arc, which would mask the ultraviolet response. Consequently it is desirable to use only those materials which are not naturally sensitive much beyond 3200 angstroms. Cadmium, uranium, and lithium have been proposed, and the first two have been used practically. In order to limit the response to the proper wavelength on the short-wavelength side, these metals can be mounted in bulbs of Corex D glass instead of quartz. If, for other purposes, it is desired to broaden the region of response up toward or into the visible blue, thorium and cerium have been suggested.

The glass bulbs of ordinary commercial cells are more or less opaque to radiation beyond about 3200 angstroms, so that, in any event, a special glass or quartz is necessary. The cesium oxide cathode, however, is very sensitive to the longer ultraviolet up to 2000 angstroms as far as measurements are available.

For certain purposes where an external battery is undesirable, cells of the barrier type are finding application, an example being as a photographic exposure meter. It is necessary, of course, to interpret the readings in terms of exposures for the various color sensitivities of emulsions by charts or suitable scales on the meter.

Table 3 lists the better-known commercial cells.

BIBLIOGRAPHY

Gudden, *Lichtelektrische Erscheinungen*, Julius Springer, Berlin, 1928.
Campbell and Ritchie, *Photoelectric Cells*, Sir Isaac Pitman and Sons, 1929.
Simon and Suhrman, *Lichtelektrische Zellen*, Julius Springer, Berlin, 1932.
Fleischer and Teichman, *Die lichtelektrische Zelle*, Theodor Steinkopff, Dresden and Leipzig, 1932.
Hughes and DuBridge, *Photoelectric Phenomena*, McGraw-Hill Book Co., 1932.
Barnard, *The Selenium Cell*, Richard R. Smith, New York, 1930.
Anderson, *Photoelectric Cells and Their Applications*, Physical and Optical Societies, London, 1930.
Geffcken, Richter, and Winckelmann, *Die lichtempfindliche Zelle als technisches Steuerorgan*, Deutsch-Literarisches Institut J. Schneider; Berlin-Tempelhof, 1933.
Zworykin and Wilson, *Photocells and Their Applications*, John Wiley & Sons, 1932.
Doty, *Selenium, A List of References, 1817–1925*, New York Public Library, 1927.
Walker and Lance, *Photoelectric Cell Applications*, Pitman, London, 1933.

Handbuch der Physik, Vol. XIX, pp. 829 ff. (Thermal cells.)
Koller, *Physics of Electron Tubes*, McGraw-Hill Book Co., New York and London, 1937.
Henney, *Electron Tubes in Industry*, McGraw-Hill Book Co., New York and London, 1937.
Reich, *Theory and Applications of Electron Tubes*, McGraw-Hill Book Co., New York and London, 1944.
Brattain and Becker, Thermistor Bolometers, *J. Optical Soc. Am.*, Vol. 36, 6, p. 354 (June 1946).
Cashman, R. J., New Photoconductive Cells, *J. Optical Soc. Am.*, Vol. 36, 6, p. 356 (June 1946).

TELEVISION PICK-UP TUBES

By V. K. Zworykin and E. G. Ramberg

8. REQUIREMENTS

The purpose of the television pick-up tube is to convert an optical image of the scene to be transmitted into an electrical signal descriptive of the light distribution in the image. In all the pick-up tubes here considered the signal is obtained by scanning in sequence a rectangular image area along a fixed number (e.g., 525) of adjoining horizontal scanning lines; with an ideal transmission system and viewing device the instantaneous signal output of the pick-up tube determines the brightness of a particular picture element (i.e., a square whose length and height are equal to the separation of two scanning lines) in the reproduced image, scanned in synchronism with the transmitted image.

A satisfactory pick-up tube must be capable of furnishing a signal that can be converted into an image with adequate detail, free from objectionable random fluctuations in brightness (noise), and faithful in geometry and tonal values over its entire area, at a reasonable illumination of the transmitted scene. Similar requirements regarding resolution, signal-to-noise ratio, uniformity, and sensitivity must also be fulfilled by the 35-mm negative film employed in commercial motion-picture production, whose properties may reasonably be taken as a standard. This is all the more appropriate since the comparison of television with motion pictures appears inescapable.

SENSITIVITY. A suitable figure of merit for the sensitivity is given by the ratio $f^2/(BA)$, where f is the f-number of the lens employed, B the brightness of the scene required to yield a good picture (in lumens per square meter), and A the area of the picture on the film (in square meters); through the factor f^2/A the figure of merit is proportional to the square of the depth of focus. Employing figures derived from motion-picture studio practice, $f = 2$, $B = 5500$ (lumens/m^2), and $A = 0.00032$ m^2 (0.5 in.^{2}), $f^2/(BA) = 2.3$.

RESOLUTION AND SIGNAL-TO-NOISE RATIO. Thirty-five-millimeter film is generally capable of resolving 1000 to 1500 lines per picture height; on the other hand, at this level the photographic grain or noise interferes seriously with the picture detail (i.e., the signal). For a ratio of the signal to the root-mean-square noise amplitude of 30–40, required to render the grain unobjectionable, the resolution must be reduced to about 500 lines. It should be noted that the root-mean-square noise amplitude employed throughout in the present discussion is only about one-sixth as great as the peak-to-peak noise amplitude, which may be observed directly on an oscilloscope screen. It is found experimentally that the signal-to-noise ratio for film remains approximately constant throughout the useful exposure range. It differs in this from the more sensitive television pick-up tubes, for which the noise is constant and the signal-to-noise ratio, hence, is lower in the low lights than in the high lights.

UNIFORMITY. The film image is geometrically faithful and uniform in response over the entire image area.

It will be seen that certain pick-up tubes exhibit higher sensitivity and signal-to-noise ratio for equal resolution than film. To this extent they enable television cameras to function more favorably than studio and news motion-picture cameras.

9. THE IMAGE DISSECTOR

The Farnsworth image dissector is shown, in schematic cross-section, in Fig. 1. At one end of the tube there is a photocathode on which a lens projects an optical image of the scene; at the other, a positive electrode with a tiny aperture, equal to a picture element in size. The magnetic field of a solenoid focuses the photoelectrons so as to form, in the plane of the aperture, a charge image of the picture on the photocathode. This charge image is swept across the aperture by the magnetic deflecting fields so that, at any instant, photoelectrons from just one picture element on the photocathode pass through the aperture. These accelerated photoelectrons fall on the first stage of an 11-stage multiplier (see article 4) built into the tube and eject a larger number of secondary electrons which are drawn through an accelerating screen to the second target electrode, leading to a further secondary-emission multiplication of the current. The output of the multiplier, finally, may be coupled by a resistance to the input of a standard video amplifier.

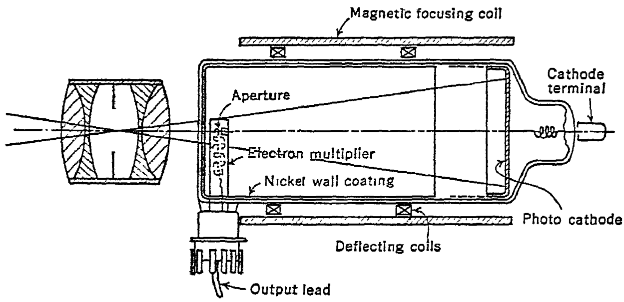


FIG. 1. The Image Dissector

The current passing through the aperture is simply the photocurrent emitted by a picture element on the photocathode. Thus, if p is the photosensitivity of the cathode in amperes per lumen, L its illumination in lux (lumen/m²), A the effective area of the photocathode in square meters, and N the number of picture elements, the signal current becomes

$$i_s = \frac{pLA}{N} \tag{1}$$

Denoting the transmitted band width by F, the shot-noise amplitude for this current is

$$\overline{i_n^2}^{1/2} = (2ei_sF)^{1/2} \tag{2}$$

Assuming $F = 5 \cdot 10^6$ sec^{-1} and a 525-line picture ($N = (525)^2 \cdot 4/3$),

$$\overline{i_n^2}^{1/2} = 2.1 \cdot 10^{-9}(pLA)^{1/2} \tag{3}$$

Hence the signal-to-noise ratio is

$$S = i_s/\overline{i_n^2}^{1/2} = 1.3 \cdot 10^3(pLA)^{1/2} \tag{4}$$

If it is assumed that $p = 20 \cdot 10^{-6}$ amp/lumen, $A = 0.01$ m² (15 in.²)

$$S = 0.58L^{1/2} \tag{5}$$

A signal-to-noise ratio of 100 would thus demand a cathode illumination $L = 29{,}000$ lux.

The multiplication provided by the multiplier should be such that the multiplied shot noise exceeds the amplifier input tube noise current, which may be estimated at $2 \cdot 10^{-9}$ amp. Since

$$\overline{i_n^2}^{1/2} = 1.6 \cdot 10^{-12}S \tag{6}$$

the multiplier gain will suffice for all recognizable picture detail ($S > 1$) if it is equal to a few thousand. The actual gain is made larger than this, reducing the required amplifier gain.

SENSITIVITY. From the above figures it follows that, in order to transmit a picture with a signal-to-noise ratio of 100, an image dissector provided with an $f/4.5$ lens would require (for an effective cathode area of 0.01 m²) a high-light brightness of the scene equal to $29{,}000\ (2 \cdot 4.5)^2 = 2.3 \cdot 10^6$ lumens/m². The figure of merit of the dissector, calculated in the same manner as for film, hence becomes

$$\frac{4.5^2}{0.01 \cdot 2.3 \cdot 10^6} = \frac{1}{1100} \tag{7}$$

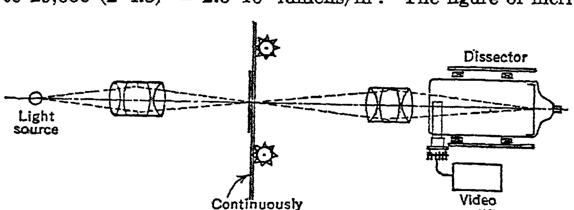


FIG. 2. Motion Picture Transmission with Image Dissector

Thus the sensitivity of the dissector is less than that of film by a factor of two or three thousand. This is a drawback for direct pick-up but is of secondary importance for motion-picture transmission. For the latter purpose, continuously moving film is generally employed (Fig. 2). The film motion provides the vertical deflection, so that only the horizontal deflection coil need be actuated. Special types of image dissectors, designed specifically for the transmission of black and white and of color film, are available. Since the standard projection speed for motion pictures is 24 frames per second and the television field frequency is 60 per second, the arrangement shown in Fig. 2 is generally modified by the insertion of an optical compensation system between the continuously moving film and the dissector tube. This causes the motion-picture frames to be scanned two and three times in alternation.

RESOLUTION. Under normal circumstances the aperture size determines the resolution of the image dissector. Change of focus with deflection is readily compensated by applying the proper correcting signals to the focusing current or the accelerating voltage in synchronism with the deflection. Fundamentally, the resolution of this type of tube is limited by the unsharpness of focus arising from the initial velocities of the photoelectrons. An increase in the number of picture elements demands, hence, either the employment of a larger photocathode, leaving the length of the tube unaltered, or a higher operating voltage and stronger focusing field.

SIGNAL-TO-NOISE RATIO. The signal-to-noise ratio of the image dissector is proportional to the square root of the scene brightness. Hence, for equal signal-to-noise ratio in the high lights, the noise will be more prominent in the low lights than for film, though less prominent than for most of the remaining pick-up tubes to be considered, for which the noise is independent of the light level. It is therefore necessary to demand a higher signal-to-noise ratio in the high lights than for film (e.g., 100 in place of 30–40). The fact that the noise becomes more noticeable in the low lights is accentuated by the circumstance that the signal output of the dissector is strictly proportional to the element brightness, i.e., that the tonal scale is not compressed by the pick-up device.

UNIFORMITY. The image dissector has excellent uniformity properties, both with regard to constancy of response over the entire picture and to the absence of geometric distortions.

10. THE ICONOSCOPE

The iconoscope, the orthicon, and the image orthicon may be classed together as storage pick-up tubes. In all of them the charge released photoelectrically from a picture element by the incident light is stored in the period intervening between two successive scannings of the element. This leads to a very great gain in sensitivity in comparison with nonstorage pick-up systems such as the image dissector.

Figure 3 shows the construction of a standard iconoscope (type 1850-A) with magnetic deflection. It is seen to consist of an electron gun whose beam is deflected across the surface of a photosensitive "mosaic" by two pairs of external deflecting coils, and the mosaic, all enclosed in a dipper-shaped envelope. An image of the scene to be transmitted is projected by a lens through an optically clear face of the envelope onto the mosaic plate whose dimensions are 4 $^3/_4$ × 3 $^9/_{16}$ in.[2]

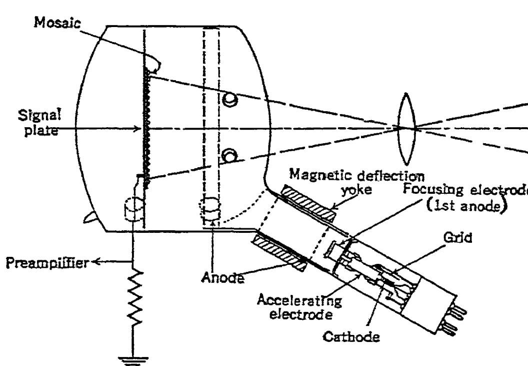


FIG. 3. The Iconoscope

A typical gun structure consists of an indirectly heated cathode enclosed in a grid cylinder with a 0.040-in. aperture, a closely spaced cylindrical accelerating electrode at full anode voltage with a defining aperture 0.002 in. in diameter, a focusing electrode or first anode, and the final anode in the form of a platinum coating on the inner wall of the gun tube. The defining aperture, which coincides approximately with the cross-over, is imaged by the equipotential electron lens formed by the two cylinders at full anode potential and the intermediate focusing electrode on the mosaic, forming a spot 0.005 in. or less in diameter; this design minimizes the current striking the focusing electrode or first anode and hence prevents disturbing secondary emission from the gun. In practice the anode is maintained at approximately 1000 volts, the focusing electrode at 300 volts, and the grid bias may be varied from −30 to −50 volts. The optimum Iconoscope beam current ranges from 0.05 to 0.2 microampere, increasing with the illumination of the mosaic.

The mosaic is a thin sheet of mica, covered, on the side facing the lens and the electron beam, with an array of minute silver globules, small compared with a picture element. These have been rendered photosensitive by a process involving oxidation, cesiation, and the subsequent evaporation of silver. On the other side the mica sheet is coated with a continuous metal film, the signal plate, which is electrically connected to the coupling resistor and the grid of the first stage of amplification. The capacitance between the signal plate and the photosensitive mosaic is of the order of 1 $\mu f/m^2$; the total capacitance between signal plate and anode coating, 10 $\mu\mu f$. The photosensitivity of the mosaic is 4–10 μa/lumen; its spectral response is shown in Fig. 4. For a mosaic illumination of 10 to 50 lux a coupling resistance of 0.1 megohm is recommended. At very low light levels (and a beam current of the order of 0.05 μa) it is proper to increase this to 1 megohm.

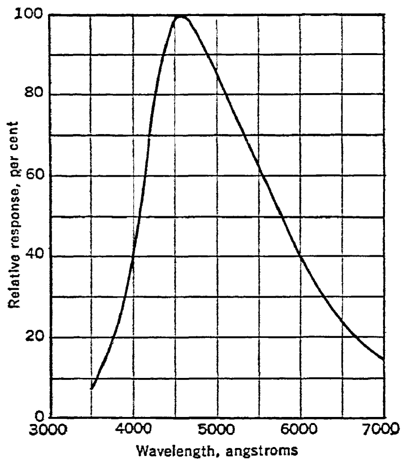


FIG. 4. Spectral Response of the Type 1850-A Iconoscope

Figure 5 shows a smaller Iconoscope (type 5527), with electrostatic deflection and transparent signal plate, which is designed primarily for industrial and amateur use. It has a 1.4-in. mosaic and operates with a beam voltage of 800 volts and a first-anode voltage between 125 and 250 volts. The cut-off voltage for the control grid is about −75 volts, and the horizontal and vertical deflection voltages are in the neighborhood of 100 volts.

OPERATION. Superficially, the operation of the Iconoscope may be described as follows: The mosaic functions as an array of minute photoelectric cells with a common anode, whose cathodes are capacitatively coupled to the signal plate. The elementary condensers so formed charge up, as the mosaic is exposed to light, by an amount proportional to the light intensity. Whenever the beam, acting as a commutator, sweeps across them, the cathode elements are returned to their equilibrium potential by collecting the requisite number of electrons from the beam, and an equal electron current passes through the signal lead. For an illuminated element on a generally dark mosaic (for which the total photocurrent would be very small), the signal current would be given by

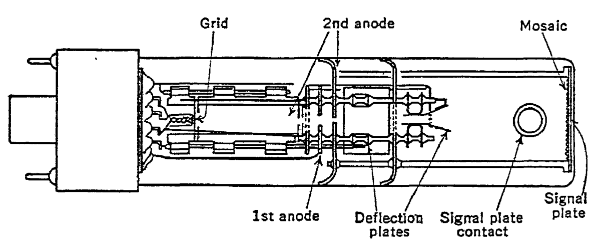


FIG. 5. Small Iconoscope with Electrostatic Deflection (Type 5527)

$$i_s = pL\frac{AT}{Nt_e} = pLA \tag{S}$$

employing the following notation:

- p photosensitivity of mosaic.
- L luminous flux/unit area of mosaic.
- A area of mosaic.
- N number of picture elements.
- T frame time.
- t_e time required to sweep over one picture element.

This would represent a gain in sensitivity relative to non-storage devices (without secondary-emission multiplication of the signal) which is equal to the number of picture elements.

The above representation is greatly oversimplified. Thus, it assumes that a mosaic scanned in darkness has a uniform potential. The actual, measured, potential distribution, at the moment when the scanning beam is approximately a third from the top of the mosaic, is as represented in Fig. 6. Figure 7 shows the variation of the potential of a particular illuminated and unilluminated element in the course of a frame time. This behavior arises in the following manner. Considering the unilluminated mosaic, an element directly under the beam emits secondary electrons whose initial kinetic energy varies from zero to a few electron volts. The secondary-emission properties of the mosaic are such that, on the average, about 4 secondary electrons are emitted for every primary electron incident from the beam. These will be able to leave the element only if the field conditions in front of it are favorable; as the element becomes, as the result of secondary emission, more positive with respect to the anode and the remainder of the mosaic, a larger proportion of the electrons will return to the element. When the element reaches a potential V_1 of the order of 3 volts positive with respect to the anode coating, only one secondary electron will leave the element for every incident beam electron and no further charging will take place. The beam current is chosen large enough to bring the picture element to the equilibrium potential V_1 in every transit.

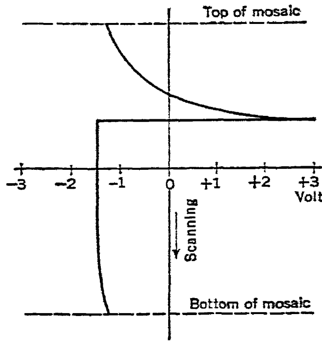


FIG. 6. Voltage Variation over Iconoscope Mosaic Scanned in Darkness

Since the mosaic is insulated, on the average only one of all the secondary electrons (and photoelectrons) which leave the element for every incident beam electron arrives at the second anode; the rest are redistributed over the remainder of the mosaic. This redistribution is influenced by the potential distribution over the mosaic (and hence, for an illuminated mosaic, also, to some extent, by the light distribution in the image) and the geometry of the tube—in particular the location of clear glass surfaces relative to the mosaic. The redistribution quickly reduces the potential of the elements immediately behind the beam and more gradually that of the more remote elements. When the elements have attained a potential V_2, of the order of $-1\ {}^1/_2$ volts, no further redistributed electrons reach them. This, thus, represents the equilibrium potential of elements not under the beam.

Consider, next, an illuminated element of the mosaic. Immediately after the beam has rendered the element about 3 volts positive with respect to the second anode, no photoelectrons are able to reach the second anode; however, an appreciable number may find their way to the elements ahead of it which have been under the beam even more recently and hence are more positive. Thus the element becomes negative less rapidly than an unilluminated element. However, the photoemission is far from saturated, a large proportion of the photocurrent returning to the element of origin. Although a larger proportion of the photocurrent will reach the anode as the element becomes more negative, the condition of incomplete saturation generally persists practically up to the succeeding passage of the scanning beam, at which point the illuminated element may have a potential V_3, a fraction of a volt above V_2. At low light levels the average photocurrent leaving a small illuminated region, with the rest of the mosaic in darkness, is approximately 20 per cent of the saturated photoemission.

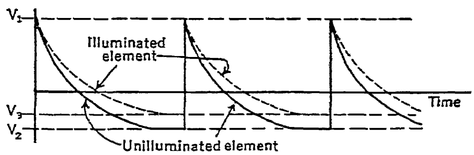


FIG. 7. Voltage Variation of an Illuminated and an Unilluminated Picture Element on the Iconoscope Mosaic

As the beam passes over the illuminated element, it returns it to the positive equilibrium potential V_1. For a small illuminated area on a dark background, the signal current is equal to the difference in the fraction of the secondary-emission current from the element which reaches the anode for the illuminated and for an unilluminated region. This is simply the current reaching the anode as an unilluminated element is raised, under the beam, from the potential V_2 to the potential V_3 (see Fig. 7). Since for both these potentials the secondary emission is, in general, saturated, and the current collected by the anode must equal the beam current, only a fourth of the stored charge can be utilized for the signal current if the secondary-emission ratio is 4. Thus the total operating efficiency at low light levels is of the order of 5 per cent (${}^1/_4$ of 20 per cent), and the signal current is given by

$$i_s = kpLA \qquad k \cong 0.05 \tag{9}$$

It may be noted that the photoelectric efficiency can be improved appreciably by illuminating slightly the photosensitive clear glass walls of the tube (backlighting), since this raises their potential.

At high light values the efficiency becomes much less; regardless of the degree of illumination, photoemission will drive an illuminated region only positive enough relative to its surroundings to prevent the departure of additional photoelectrons. Thus the Iconoscope signal is compressed in the high lights. The preferred collection of the redistributed electrons by the more positive areas of the mosaic enhances this effect. At very high light levels the Iconoscope signal is determined by the photoelectric charge stored during the line scan of the beam preceding the scanning of the picture element considered: Since under the beam an element becomes positive by 3 volts relative to the second anode, the photoemission of the neighboring element on the line ahead of the scanned line is saturated even if it is positive by several volts relative to its other neighbors ("line sensitivity"); an opposing voltage of 1–2 volts suffices to suppress the photoemission.

The signal output characteristic of the 1850-A Iconoscope is shown in Fig. 8.

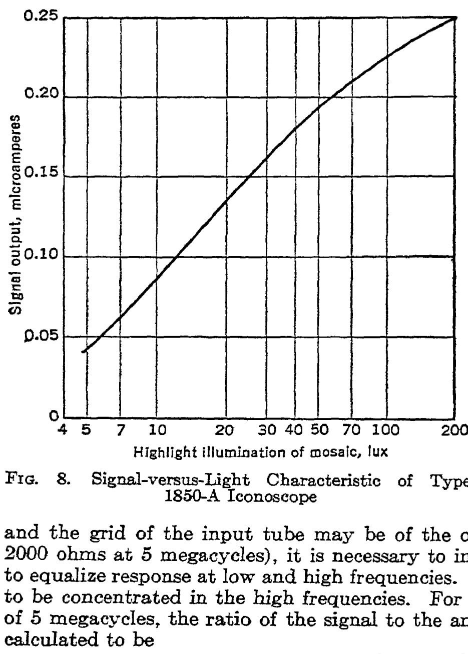


FIG. 8. Signal-versus-Light Characteristic of Type 1850-A Iconoscope

SIGNAL-TO-NOISE RATIO. The principal source of noise in the conventional Iconoscope pick-up system is the noise introduced by the input tube of the signal amplifier, which may be represented as thermal noise from an equivalent resistor r_T added to the resistive component of the coupling network between the Iconoscope and the amplifier. Since, for coupling resistances of the order of 0.1 to 1 megohm, the coupling impedance is primarily capacitative for the high frequencies of the video band (the combined capacitance to ground of the signal plate and the grid of the input tube may be of the order of 15 $\mu\mu$f, corresponding to about 2000 ohms at 5 megacycles), it is necessary to insert a peaking network in the amplifier to equalize response at low and high frequencies. This network causes the noise spectrum to be concentrated in the high frequencies. For low-noise input tubes and a video band of 5 megacycles, the ratio of the signal to the amplitude of the integrated noise may be calculated to be

$$S = 5 \cdot 10^8 i_s \tag{10}$$

i_s being measured in amperes. It should be noted that about 2 or 3 times as much of this peaked noise can be tolerated by the observer as shot noise (e.g., from a multiplier) distributed uniformly over the spectrum.

SENSITIVITY. In practice a scene brightness from 6000 to 17,000 lumen/m² is found to give satisfactory pictures with an 1850-A Iconoscope used in conjunction with a lens of f/5.6 or smaller aperture. The target area of the mosaic is approximately 0.011 m² (17 in.²). Employing the lowest values both for the illumination and the f-number, the figure of merit used as a measure of the sensitivity becomes

$$\frac{5.6^2}{6000 \cdot 0.011} = 0.48 \tag{11}$$

This is about $^1/_5$ that for film; the formula for the signal-to-noise ratio yields a value of 100, in harmony with direct measurements. This is greater than for film by a factor between 2 and 3. Such a factor is needed to render the noise unobjectionable in the low lights, since the noise does not decrease in proportion with the signal. The fact that the signal output of the Iconoscope has less contrast than the original image, i.e.,

$$\frac{\Delta i_s}{i_s} = \frac{1}{k}\frac{\Delta L}{L} \quad \text{where } k = 2 - 3 \tag{12}$$

a condition which is generally compensated in the viewing tube, leads to a useful reduction in the difference between the signal-to-noise ratios for the high lights and the low lights.

RESOLUTION. The resolution of the Iconoscope, as for the other pick-up tubes here considered, may be extended to better than 1000 lines; for the 525-line standard it is customary to peak the high-frequency response electrically so as to keep the response level up to 500 lines.

UNIFORMITY. The description of the operation given above makes it evident that the signal output for any picture element is not simply related to its brightness; the redistribution of the secondary and photoelectrons makes it dependent both on the geometrical position of the element and on the light distribution in the remainder of the picture. Hence spurious signals—"shading"—are introduced and must be compensated electrically with the aid of shading controls. They become particularly troublesome when large dark areas are present in the scene.

11. THE MONOSCOPE

The monoscope is not, strictly speaking, a pick-up tube. It merely serves to supply a standard picture signal, whose character is prescribed by the preparation of the target electrode. As such it has found application primarily in the testing and aligning of the components of a television system other than the pick-up tube itself.

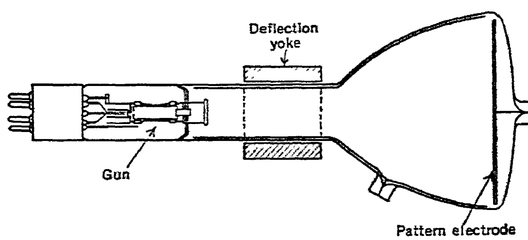


FIG. 9. The Monoscope

Figure 9 shows the construction of a type 2F21 monoscope. It consists of an Iconoscope gun and a target plate (3 1/16 by 2 5/16 in.), normal to the axis of the gun, mounted in a pear-shaped envelope. The target plate consists of sheet aluminum on which a pattern has been printed with a carbon ink (Fig. 10). The magnetic beam deflection and the circuit connections are the same as for an Iconoscope, with the distinction that the anode coating is maintained at a potential 50–200 volts positive with respect to the target plate.

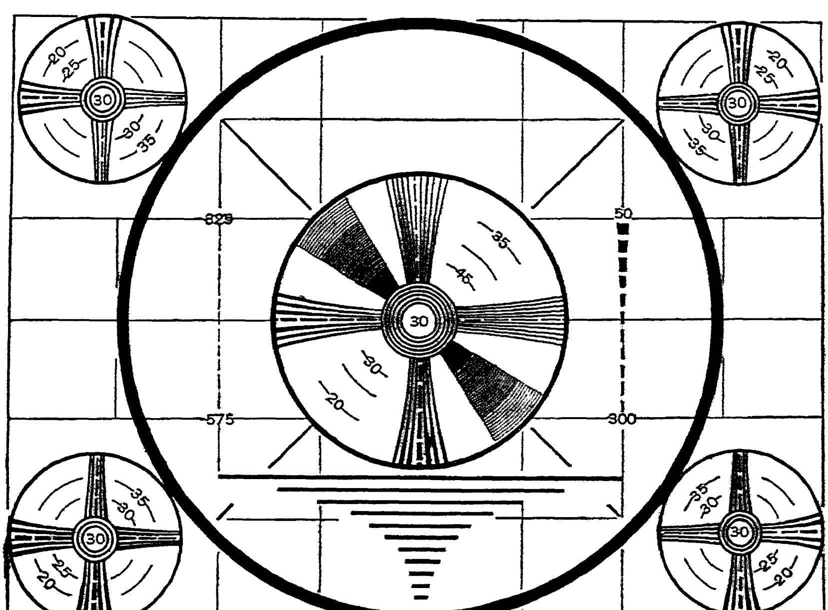


FIG. 10. Pattern on the Type 2F21 Monoscope

The operation of the monoscope depends on the much greater secondary emission ratio of the slightly oxidized aluminum (~3) as compared with that of carbon (<1). Whenever the beam strikes the aluminum, a signal current approximately twice as great as the beam current flows through the signal lead in such a direction as to make the grid of the input tube positive; when it strikes the carbon, a small current in the opposite direction tends to make it negative. If a monoscope replaces an Iconoscope without changes in the amplifier, the portions of the pattern printed in carbon ink will appear bright in the final image. Dark and light may, of course, be interchanged by adding or subtracting one stage in the video amplifier. A peak-to-peak signal amplitude of several microamperes may be obtained with this tube.

12. THE ORTHICON

The primary defects of the Iconoscope, namely, shading and low efficiency of operation, both arise from the redistribution of secondary electrons and photoelectrons on the mosaic. This, in turn, is a consequence of the fact that the equilibrium potential of the mosaic under the beam is close to anode potential, in fact, slightly positive with respect to it. If, however, the beam arriving at the mosaic has, initially, a kinetic energy of 10 electron volts or less, the secondary emission ratio is less than unity and the potential of the mosaic will drop to a value slightly below that of the emitting cathode; at this equilibrium potential no additional electrons can reach the mosaic. Instead, the beam electrons reverse their direction at a point close to the mosaic and are collected by some electrode at positive potential. Portions of the mosaic which are in complete darkness remain continuously at the equilibrium potential and, hence, give rise to no signal current. Illuminated areas, on the other hand, lose electrons, between successive scannings, in exact proportion to the quantity of light incident on them. The low-velocity beam, as it sweeps over such areas, supplies just enough electrons to the mosaic to neutralize the stored charge and causes the passage of an equal signal current through the signal lead. In brief, the actual operation of such a low-velocity Iconoscope fits perfectly the original, oversimplified version given for the operation of the Iconoscope.

The practical realization of the low-velocity Iconoscope demands fulfilment of two conditions: (1) in order that the equilibrium potential (and scanning spot) may be uniform over the mosaic, the beam must be perpendicular to the mosaic at all points; (2) to keep the effective spot size small, the lateral velocity components of the electrons must be kept small. These requirements are met by the special methods of beam deflection and focusing incorporated in the *orthicon.*

The envelope of a typical orthicon (Fig. 11) is a 4-in.-diameter tube 14 in. long with a short neck for the gun at one end and a flat, clear window for the transmission of the

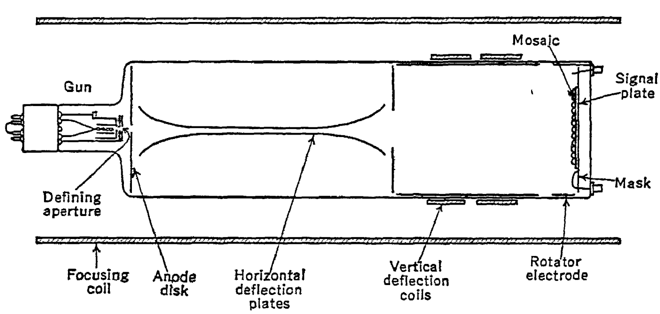


FIG. 11. Orthicon with Electrostatic Horizontal Deflection

optical image at the other. It is inserted in a long solenoid providing a uniform longitudinal magnetic field of 0.007 weber/m^2. The envelope contains, in addition to the gun, a pair of curved electrostatic deflection plates which occupy half of the tube nearest the gun, and a mosaic with a translucent signal plate having a target area 2 $^5/_{16}$ by 1 $^3/_4$ in. The horizontal deflection is accomplished with the aid of the electrostatic deflection plates; the vertical deflection, by a pair of coils mounted over a portion of the second half of the 4-in. cylinder.

The principal peculiarities of the deflection and focusing properties of the orthicon are a consequence of the presence of the longitudinal magnetic focusing field. In particular, near the mosaic, where the magnetic field lines are perpendicular throughout to the mosaic surface, the field assures the normal incidence of the electron beam: In a strong magnetic field the electrons spiral about the magnetic field lines.

Proceeding from one end of the tube to the other, the electrons leaving an indirectly heated cathode through an aperture in the grid cylinder (cut-off voltage, −40 volts) are accelerated through an aperture at 225 volts and restricted to a narrow pencil by a defining aperture 0.0025 in. in diameter which is electrically connected to the accelerating aperture. This pencil, the scanning beam, is focused by the longitudinal field, which is strong enough to form a succession of images of the defining aperture less than 2 in. apart. Its current is generally of the order of 0.3 microampere, i.e., just enough to discharge the most strongly illuminated portions of the mosaic. After passing through another larger aperture in the anode disk (at 250 volts), which separates the gun chamber from the deflection chamber, the electrons enter the electrostatic deflecting field between flared plates. The simultaneous action of the electrostatic field and the longitudinal magnetic field causes a lateral displacement of the beam parallel to the plates and proportional to the deflection voltage (160 volts, peak to peak). The flaring, making the increase and decline of the deflecting field gradual, prevents the development of cycloidal loops in the crossed fields and, hence, the acquisition of considerable lateral velocity components by the beam electrons.

Next, the beam passes through the magnetic deflecting field (0.0025 weber/m^2, peak to peak) which simply warps the field lines, so that the beam experiences a second displacement, in the direction of the deflecting field (*not* at right angles thereto). After leaving this deflecting field the electrons pass through a decelerating ring electrode at 100 volts (designated as "rotator electrode," since the simultaneous action of the lateral components of the decelerating field and the longitudinal magnetic field causes a slight rotation of the scanning pattern) painted on the envelope to the mosaic, which is inserted in a mask maintained 3 volts negative with respect to the cathode. Both the signal plate and the photosensitive mosaic are translucent. Although the requirement of translucence reduces the photoemission of the mosaic, as compared with that of the Iconoscope, this is more than compensated by the greater efficiency of operation.

SENSITIVITY. It is found in practice that a studio scene with a brightness of 700 lumens/m^2, transmitted with an $f/2$ lens, will yield a picture with a signal-to-noise ratio of 100. The target area being 0.0026 m^2 (4 in.^{2}), the figure of merit of the orthicon becomes

$$\frac{2^2}{700 \cdot 0.0026} = 2.2 \tag{13}$$

This is the same as the figure for film and better by a factor of 5 than that for the Iconoscope. However, since the orthicon does not compress the brightness scale in the same manner as the Iconoscope, but has a strictly linear response throughout, a signal-to-noise ratio higher by a factor of 2 or 3 may be required in the transmission of naturally contrasty outdoor scenes to attain an equal freedom from noise in the low lights. Hence, under such circumstances a more appropriate value for the figure of merit is 1.

SIGNAL-TO-NOISE RATIO AND RESOLUTION. A signal-to-noise ratio of 100 is readily obtained. Attempts to exceed this value by increasing the brightness of the light image frequently result in a loss of resolution and a local distortion of the scanning pattern at the boundaries between bright and dark areas.

UNIFORMITY. If the image brightness is kept within the normal operating range the signal output yields a faithful representation of the geometry and tonal values of the scene. Scene details of excessive brightness (e.g., the explosion of flash bulbs) may, however, cause a portion of the scene to be blacked out. At such points the photoemission charges the mosaic up to a positive potential at which the secondary-emission ratio of the beam electrons exceeds unity, so that the beam renders the illuminated area more positive instead of discharging it. After the cause has been removed, normal operating conditions are gradually re-established by surface leakage.

13. THE IMAGE ORTHICON

In the image orthicon a very great gain in sensitivity has been combined with the freedom from spurious signals at low light levels which is characteristic of the orthicon and the stability of operation at high light levels characteristic of the Iconoscope at the expense of greater complexity of construction and alignment. The principal features which distinguish the image orthicon from the ordinary orthicon are (1) an electron-optical imaging section, making possible the employment of a more sensitive, continuous, photocathode and secondary-emission multiplication at the target; (2) a two-sided target, permitting limitation of the target voltage to a value sufficiently low to insure stability at all light levels by providing a separate collector on the side opposite to the scanned side; and (3) a secondary-emission signal multiplier, for the return beam current, which renders the signal output sufficiently large that the shot noise in the beam, rather than amplifier noise, determines the noise content in the reproduced picture.

Figure 12 shows, schematically, the construction of the type 2P23 image orthicon. The tube has approximately the same length, but, with the exception of the short end section, a much smaller diameter than the 1840 orthicon. As with the latter tube, a long focusing solenoid must be provided which envelops all the tube except the gun and multiplier portion at the right extremity. The light image is projected on the flat transparent photocathode (maintained at −300 volts) at the left end of the tube, and the photoelectrons released by the light are focused by the magnetic field through a very fine-mesh (500–1000 meshes per inch), high-transmission screen on the target. Since the secondary-emission ratio of the glass target screen for 300-volt electrons is much greater than unity, a positive charge pattern corresponding to the light distribution in the scene is formed on the target. The secondary electrons are drawn to the closely spaced target screen, which is usually maintained at a potential slightly positive (e.g., at +1 volt) with respect to the scanning-beam cathode.

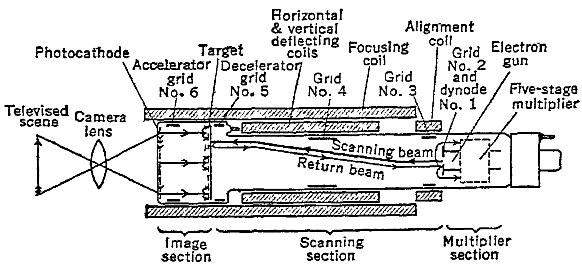


FIG. 12. The Image Orthicon

The target itself is a very thin disk of low-resistivity glass. Its properties are such that in the course of a frame time potential differences between the two sides of the target built up at the instant of scanning are neutralized by conduction while the transverse leakage of stored charge between neighboring picture elements still remains negligible.

The gun of the image orthicon has the same general construction as that of the orthicon. However, the defining aperture is formed in the final disk electrode, which is exposed to the return beam. The beam deflection is magnetic throughout, since this causes the return beam to travel practically the same path as the scanning beam in reverse direction, striking finally the electrode containing the defining aperture. The return beam itself carries the signal; when the scanning beam strikes a portion of the target which has been rendered positive by secondary emission, electrons equal in number to those lost in the course of a frame time by the element under consideration are abstracted from the beam, reducing, correspondingly, the current in the return beam.

The return beam strikes the defining-aperture disk with an energy of the order of 200 electron volts and ejects from it secondary electrons which, persuaded by the lower potential of the electrodes facing it and the higher potential of a second stage of a pinwheel multiplier structure surrounding the gun, spill over into the same emitting a larger number of electrons, which are drawn to the next stage. The total gain of the 1500-volt five-stage multiplier is from 200 to 500, which is adequate to raise the shot noise level in the beam ($i_b^{1/2} \cdot 10^{-6}$ amp) above the amplifier input tube noise current ($2 \cdot 10^{-9}$ amp, both for a 5-megacycle band width). At very low lights the proper value of the beam current may, under ideal circumstances, be as low as 10^{-10} amp, leading to a useful multiplier gain of 200; for a high-light picture a gain of 20 would suffice.

SIGNAL VERSUS LIGHT CHARACTERISTICS. Figure 13 shows a typical variation of the signal output of the image orthicon with the high-light illumination of the photocathode. The photosensitivity of the photocathode is of the order of 10 microamperes per lumen.

In the low-light range the image orthicon functions just as the orthicon, the signal current being proportional to the light signal. At the knee of the curve the secondary-emission charges the target just to the potential of the target screen; beyond this point the response curve flattens out, since an increasing number of secondary electrons are forced to return to the emitting element. As the target becomes sufficiently positive to lose only as many electrons by secondary emission as it receives from the photocathode, the response curve becomes completely flat. This does not mean, however, than no intensity differences are transmitted.

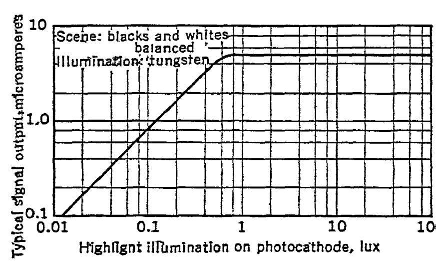


FIG. 13. Signal-versus-Light Characteristic of Type 2P23 Image Orthicon

Redistribution of secondary electrons near boundaries between areas of different intensity of bombardment (different brightness) results in potential differences at such boundaries. Thus a bright spot on a less bright background is transmitted as a bright spot with a dark halo on a bright background.

SENSITIVITY. It is found that an image orthicon provided with an $f/2$ lens and capable of transmitting a picture with a signal-to-noise ratio of 100 can do so if the scene brightness is 20 lumens/m^2. Since the target area is approximately 0.0008 m^2 (1.2 in.^{2}), the figure of merit of the image orthicon becomes

$$\frac{2^2}{20 \cdot 0.0008} = 250 \tag{14}$$

This is approximately 100 times as great as the figure for film and for the ordinary orthicon and over 500 times as great as that for the Iconoscope. A factor of 5 in this gain must be attributed to the increased photosensitivity of the photocathode and the secondary-emission amplification at the target; the remainder, to the signal multiplication in the multiplier. To maintain freedom from objectionable noise in the low lights of high-contrast scenes, it may be necessary, just as with the orthicon, to increase the signal by a factor of 2 or 3, reducing the figure of merit to unity. It may be noted that the sensitivity of the tube is sufficient to transmit pictures with some entertainment value even at a scene brightness of 0.2 lumen/m^2, corresponding to the brightness of light objects in full moonlight.

RESOLUTION. The resolution of the image orthicon may be limited by the electron-optical imaging process, the target screen, transverse leakage on the glass target, and scanning spot size, the last being influenced by the defining aperture, the angle of approach of the beam to the target, the initial velocity distribution of the beam electrons, and the potential of the scanned area. In practice it is possible to attain a resolution of 500 lines as with the other tubes.

UNIFORMITY. In the low-light range (the sloping part of the curve in Fig. 13), the signal output is a linear function of brightness. For higher light values the tonal scale is compressed, and, ultimately, contrasts, rather than absolute light values, are transmitted primarily. This condition does not detract materially, however, from the apparent naturalness of most reproduced pictures. Geometric distortions are inappreciable, although slight non-uniformities in the target and the presence of the target screen tend to make the picture somewhat inferior to that transmitted by the other pick-up tubes described.

14. FIELDS OF APPLICATION OF PICK-UP TUBES

The characteristics of the several pick-up tubes here discussed mark out spheres of application for which each is particularly suitable. Thus, the relatively insensitive image dissector, with its freedom from signal distortion, may be employed for the transmission of motion pictures, for which very high light levels can readily be provided. The standard Iconoscope is well suited for both movie and studio work, where the light distribution can be controlled so as to simplify the compensation of shading. Its smaller, 2-in. version is a convenient television pick-up device for industrial and experimental purposes. Spot pick-up, with the attendant unpredictable conditions of lighting, demands the employment of the image orthicon which, in view of its greater complexity and somewhat inferior picture quality, may under other circumstances be replaced advantageously by the less sensitive tubes. It is to be hoped and expected that further development of the image orthicon, as the most recent of the pick-up devices, will raise the level of its picture quality to that of the older pick-up tubes.

LUMINESCENT AND TENEBRESCENT MATERIALS

By H. W. Leverenz

Luminescence is a production of light in excess of thermal radiation (see ref. 5 on p. 15-41). Thermal radiation is emitted by electrons, atoms, ions, and molecules oscillating or rotating singly or in groups as occasioned by thermal agitation.

An ideal thermal radiator is the perfect *black body*, which has complete absorptivity at all wavelengths; i.e., it has oscillators available at all frequencies. The monochromatic emissive power, E_ν, of a perfect black body (in vacuum) at frequency ν (in sec^{-1}), is a function of temperature, T (in degrees Kelvin), according to Planck's radiation law:

$$E_\nu = 4.63 \times 10^{-50}\nu^3(\varepsilon^{h\nu/kT} - 1)^{-1} \quad \text{watts/m}^2 \tag{1}$$

where $h = 6.624 \times 10^{-34}$ joule-sec (Planck's constant).
$\varepsilon = 2.71828\cdots$ (base of Napierian logarithms).
$k = 1.38 \times 10^{-23}$ joule/deg (Boltzmann's constant).

The peak wavelength, λ_{max}, of the broad emission band of black-body radiation varies with absolute temperature according to Wien's displacement law:

$$\lambda_{max} = 2.897 \times 10^{-3}T^{-1} \text{ m } (1 \text{ m} = 10^6 \text{ microns } (\mu) = 10^{10} \text{ angstrom units } (\text{Å})) \tag{2}$$

The total emissive power, E_T, of a black body is proportional to the fourth power of the absolute temperature according to the Stefan-Boltzmann law:

$$E_T = 5.67 \times 10^{-8}T^4 \quad \text{watts/m}^2 \tag{3}$$

At room temperature, $\approx$300 deg Kelvin, λ_{max} is in the far infrared at 9.7 microns, and E_T is only 459 watts/m². It should be noted that the nature of the material plays no role in eqs. (1), (2), and (3). Although thermal radiation is emitted by all materials at temperatures greater than 0 deg Kelvin such radiation does not become visible until the temperature is raised above about 1000 deg Kelvin, when incandescence is observed. A temperature of the order of 6500 deg Kelvin is required to shift the peak wavelength, λ_{max} of thermal radiation into the visible region of the spectrum (4000 to 7000 Å). Thus far, no solid material has been developed to endure prolonged operation above 4000 deg Kelvin, and so most of the energy emitted from solid incandescent materials lies in the infrared and their efficiencies of light production are generally less than about 5 per cent.

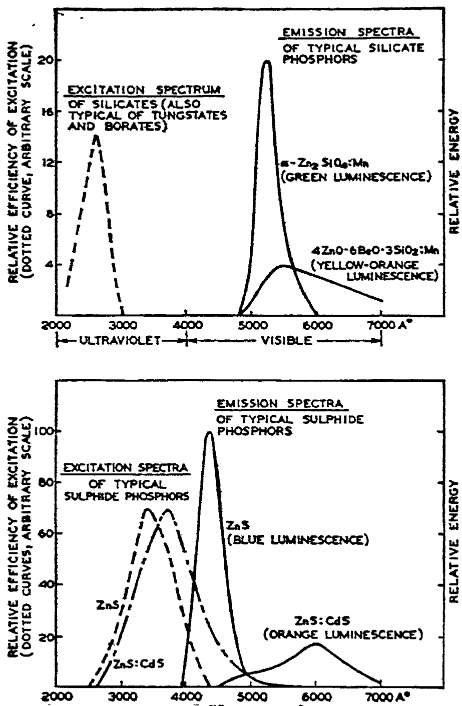


FIG. 1. Excitation and Emission Spectra of Some Typical Phosphors

Luminescence is occasioned by absorbed photons, so-called undulatory energy (e.g., ultraviolet, x-rays, γ-rays), or corpuscular energy (e.g., cathode rays or α-particles) which excite electronic transitions directly rather than through the intermediate stage of thermal agitation of atoms and ions. Luminescence emission is usually in the form of spectral lines or narrow bands superimposed on the broad band of thermal radiation from a material. The spectral distributions and efficiencies of luminescent materials are determined largely by their chemical compositions and, if the materials are solids, by their crystalline structures. The characteristic monochromatic spectra of attenuated gases are relatively simple luminescences whose efficiencies may approach 100 per cent for the case of resonance radiation ($\lambda_{excitation} = \lambda_{emission}$). In liquids and solids, however, the perturbations imposed by near neighbors of a luminescing atom or ion complicate the mechanism and generally lower the efficiency of luminescence.

The generic term luminescence is commonly modified by a prefix indicative of the excitant used to cause luminescence. For example, *photoluminescence* is luminescence excited by photons, and *cathodoluminescence* is luminescence excited by cathode rays. A further distinction is made with respect to duration of luminescence after cessation of excitation; i.e., *fluorescence* lasts less than about 10^{-8} second whereas *phosphorescence* lasts longer than about 10^{-8} second. The value of 10^{-8} second is the approximate lifetime of excited non-metastable isolated atoms or ions and serves as an arbitrary demarkation between fluorescence and phosphorescence. Materials, such as gases, liquids, organic materials, and many glasses which exhibit fluorescence are called *fluors*; while phosphores-

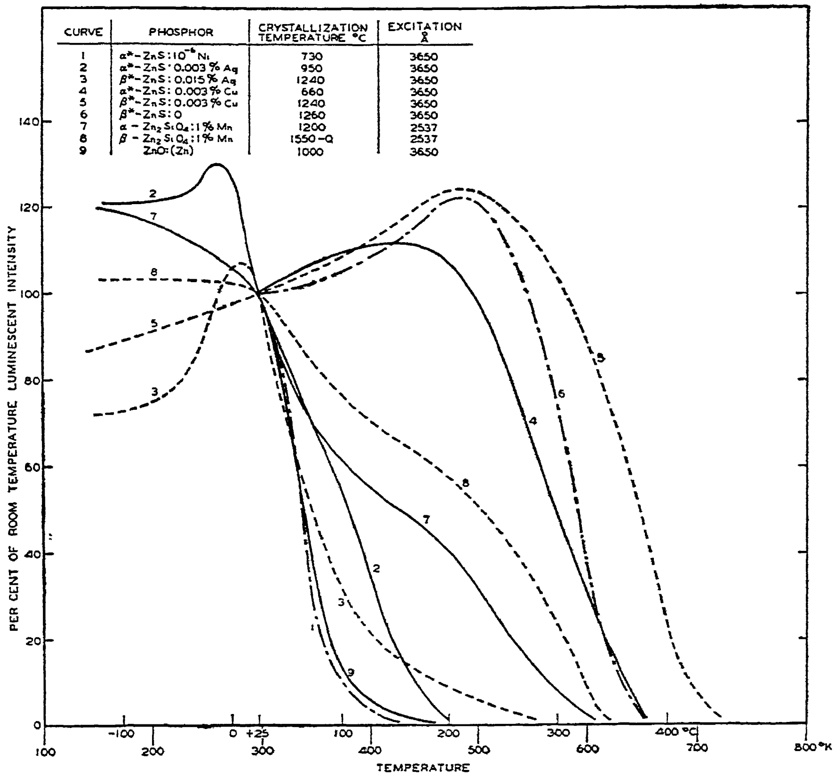


Fig. 2. Photoluminescences of Some Phosphors as a Function of Temperature

cent materials, which are chiefly crystalline inorganic materials, are called *phosphors.* Some typical excitation and emission spectra of phosphors are shown in Fig. 1; several temperature-dependence curves of phosphor photoluminescences are shown in Fig. 2, and some typical excitation and decay characteristics of phosphors are indicated in Fig. 3.

Phosphor light outputs may be modulated by three methods:

1. Positive modulation of luminescence is the normal increase of light output with increasing excitation density at temperatures below the fairly critical temperature, T_c, above which the efficiency of luminescence sharply decreases.

2. Negative modulation of luminescence is accomplished by increasing the temperature of an excited phosphor above T_c and thereby decreasing the luminescence.

3. Positive modulation of incandescence is accomplished by further raising the temperature of the phosphor until incandescence supplants luminescence.

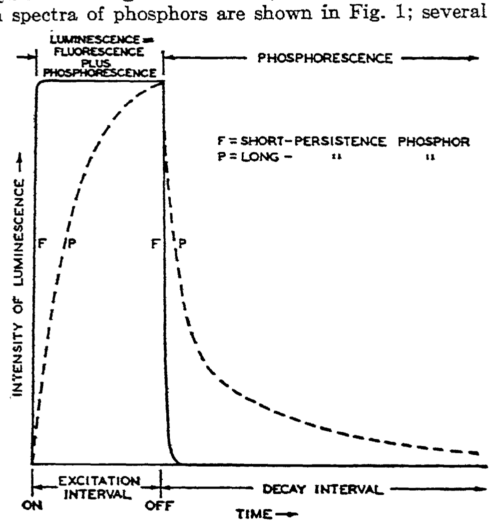


Fig. 3. The Relationships of Luminescence, Fluorescence, and Phosphorescence

The last two methods of modulation involve thermal inertia of matter as contrasted with the purely electronic transitions in positive modulation of luminescence. Positive modulation of luminescence is unique in allowing useful modulation up to frequencies of the order of 10^7 cycles per second, the limit for any particular phosphor being inversely proportional to its characteristic decay time (the time taken to decay to an arbitrary percentage, e.g., 1 per cent, of the luminescence at the last instant of excitation).

Tenebrescence is any non-intrinsic absorption of light induced in a material. For example, normally colorless potassium chloride, KCl, whose intrinsic absorption is in the far ultraviolet (left side of Fig. 4), may have an absorption band induced in the visible part of the spectrum by irradiation with cathode rays or x-rays (see right side of Fig. 4). The induced darkening (tenebrescence) may be bleached by irradiating the darkened material with light having wavelengths lying within the induced absorption band. The bleaching is accelerated by heat. Tenebrescent materials become increasingly difficult to bleach as the duration and intensity of the primary irradiations used to induce tenebrescence are increased. The relatively unbleachable absorptions are similar to those of pigments or dyes which convert absorbed photons into heat. Bleachable tenebrescences are ascribed to temporary trapping of electrons; unbleachable tenebrescences apparently involve concomitant ionic displacements.

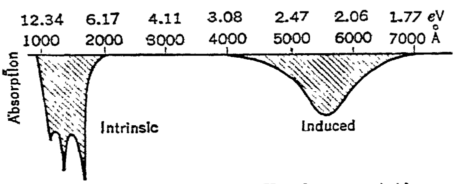


FIG. 4. Intrinsic and Induced (Tenebrescence) Absorption Bands of Potassium Chloride

Tenebrescent materials, such as the crystalline halides of alkali or alkaline-earth metals, are called *scotophors*.

15. PREPARATION AND NOTATION OF PHOSPHORS

Successful preparations of synthetic phosphors require highly specialized chemical and physical operations wherein even the most skilled and careful workers sometimes have difficulty in reproducing results. Phosphor ingredients must be purified to contain less

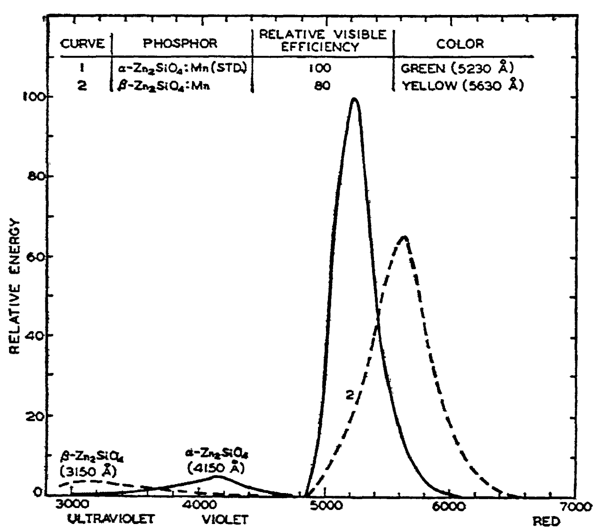


FIG. 5. Cathodoluminescence Spectra of α- and β-Zn_2SiO_4:Mn with and without Manganese Activator

than about 10^{-4} per cent of undesirable metallic-ion impurities (e.g., iron, nickel, and chromium), since as little as 10^{-4} per cent of combined nickel in a zinc-cadmium-sulfide phosphor lowers efficiency about 25 per cent. On the other hand, 10^{-4} per cent of combined silver in the foregoing *pure* phosphor increases efficiency 100 per cent. The final crucial step in preparing phosphors is crystallization, where the purified ingredients are

CURVE	PHOSPHOR	TEMP. FOR 2hr.	RELATIVE VISUAL RESPONSE	NATURAL COLOR	COLOR OF LUMINESCENCE
1	ZnS:0.008% Ag	940°	18.5(1a)	WHITE	LIGHT BLUE
2	ZnS(80)CdS(20):0.01%Ag	"	27.0	LIGHT GREEN WHITE	VERY LIGHT BLUE GREEN
3	ZnS(60)CdS(40): "	"	66.3	VERY LIGHT GREEN	VERY LIGHT CREAM GREEN
4	ZnS(50)CdS(50): "	"	100.0(4a)	LIGHT YELLOW	LIGHT GREEN YELLOW
5	ZnS(40)CdS(60): "	"	63.7	LIGHT CREAM YELLOW	LIGHT YELLOW ORANGE
6	ZnS(20)CdS(80): "	"	9.4(6a)	TAN ORANGE	LIGHT RED
7	CdS:0.02%Ag	"	—	LIGHT BROWN ORANGE	RED

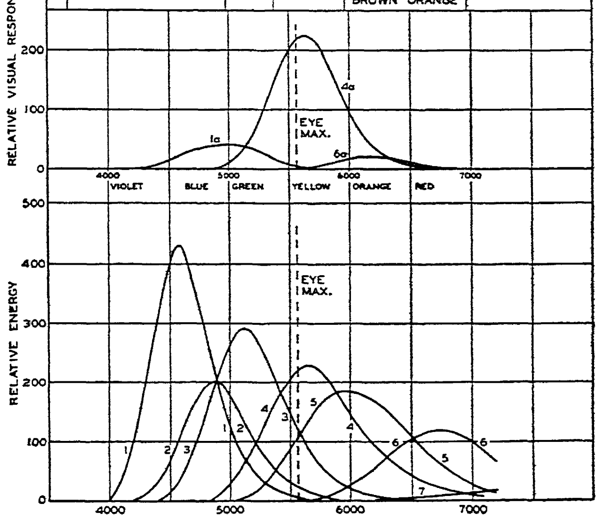


FIG. 6. Cathodoluminescence Emission Spectra of Some Silver-activated Zinc-cadmium-sulfide Phosphors. Upper curves are the relative visual response characteristics of Nos. 1, 4, and 6.

mixed in fused-silica or platinum crucibles and heated in electric resistance furnaces, generally to temperatures between 600 and 1600 deg cent. The resultant phosphors are masses of tiny crystals ranging from less than 0.01 to about 100 microns in diameter. Most phosphor crystals average about 1 to 15 microns in diameter. Some typical initial compositions and corresponding notations of the resultant phosphors are given in Table 1.

The luminescence emission spectra of phosphors are strongly influenced by changes in crystallization and composition, as shown in Figs. 5, 6, and 7. Phosphors such as P3, P4(Y), and P7/2 belong to "families" wherein gradual base-material variations enable one to produce emission spectra which may be varied continuously from one end of the visible spectrum to the other. Other properties, such as

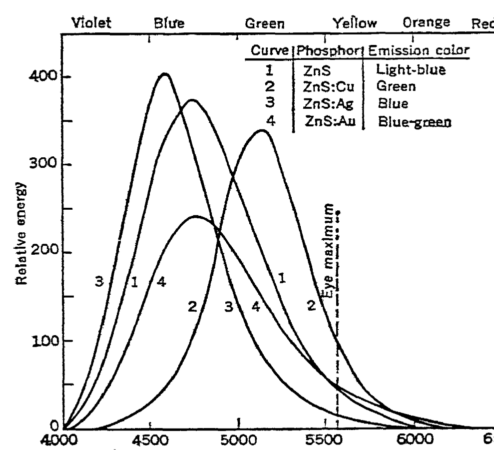


FIG. 7. Cathodoluminescence Emission Spectra of Zinc-sulfide Phosphors Prepared with (1) no added activator, (2) copper activator, (3) silver activator, and (4) gold activator

absorption spectrum, efficiency, and phosphorescence, are also considerably affected by changes in the structures and compositions of phosphors.

Table 1. *Approximate* **Compositions and Notations of Some Useful Phosphors**

RMA Code	Base Material Ingredients, grams	Activator Salt, grams	Flux, grams	Crystallization Temperature, deg cent	Phosphor Notation
P1	81 ZnO + 31 SiO_2	0.5 MnO	....	1250	α-Zn_2SiO_4:Mn
P2	100 ZnS	0.02 $CuCl_2$ + 0.04 AgCl †	6 NaCl	1250	β *-ZnS:Ag †:Cu
P3 P4(Y)	65 ZnO + 2.5 BeO + 31 SiO_2	0.4 MnO	....	1250	8ZnO·BeO·5SiO_2:Mn
P4(B) P6(B) P11	100 ZnS	0.002 to 0.02 AgCl	2 NaCl	950	α *-ZnS:Ag
P4(Y)	48 ZnS + 52 CdS	0.02 AgCl	2 NaCl	950	ZnS(48)·CdS:Ag
P5	57 CaO + 232 WO_3	....	....	1000	$CaWO_4$:[W]
P6(G)	60 ZnS + 40 CdS	0.02 AgCl	2 NaCl	950	ZnS(60)·CdS:Ag
P6(R)	38 ZnS + 62 CdS	0.02 AgCl	2 NaCl	950	ZnS(38)·CdS:Ag
P7/1 P14/1	100 ZnS	0.02 AgCl	4 NaCl + 2 $BaCl_2$	1250	β *-ZnS:Ag
P7/2	86 ZnS + 14 CdS	0.02 $CuCl_2$	2 NaCl	1250	ZnS(86)·CdS:Cu
P12	103 ZnF_2	0.5 MnF_2	....	1000	ZnF_2:Mn
P14/2	75 ZnS + 25 CdS	0.01 $CuCl_2$	2 NaCl	1200	ZnS(75)·CdS:Cu
P15	100 ZnO	(Heat in CO or H_2 at 1000° C)			ZnO:[Zn]
....	100 SrS (or SrS + SrSe)	0.03 $SmCl_3$ (or $TbCl_3$) + 0.03 $Eu_2(SO_4)_3$ (or $CeCl_3$)	6 (CaF_2 + $SrSO_3$)	1000	SrS:SrSe:Sm(Tb):Eu(Ce)

α * = cubic; β * = hexagonal.
† = optional. (B) = blue, (G) = green, (Y) = yellow, (R) = red. The Ag and Cu activator ions may also be added as nitrates.

16. MECHANISMS OF PHOSPHORS

Energy transductions during excitation and emission of phosphor luminescences have the following chronological sequence:

$$E_{\text{primary}} = aE_{\text{reflected}} + bE_{\text{absorbed}} + cE_{\text{escaped}} \tag{4}$$

$$bE_{\text{absorbed}} = b_1E_{\text{transmitted to activator centers}} + (b - b_1)E_{\text{heat}} \tag{5}$$

$$b_1E_{\text{in activator centers}} = b_2E_{\text{stored}} + (b_1 - b_2)E_{\text{internal fluorescence}} \tag{6}$$

$$b_2E_{\text{stored}} = b_3E_{\text{internal phosphorescence}} + (b_2 - b_3)E_{\text{heat}} \tag{7}$$

$$(b_1 - b_2)E_{\text{internal fluorescence}} + b_3E_{\text{internal phosphorescence}} = b_4E_{\text{external luminescence}} + (b_1 + b_3 - b_2 - b_4)E_{\text{heat}} \tag{8}$$

where $a + b + c = 1$; $1 \geq b \geq b_1 \geq b_2 \geq b_3 \geq b_4$: and E_{escaped} is the residual primary energy which completely penetrates the phosphor crystals or which emerges from the side of incidence owing to internal scattering.

Luminescence emission is occasioned when a bound electron in energy state E_0 is excited to a higher allowed energy state, E_{ex}, and returns to the same or an intermediate energy level, E_{act}, emitting the energy difference, $\Delta E = E_{\text{ex}} - E_{\text{act}}$, as a photon of light. The relations between energy ΔE (in joules), frequency ν (in cycles per second), and wavelength λ (in meters), of photons are given by:

$$\Delta E = h\nu = \frac{hc}{\lambda} \tag{9}$$

where $c = 3 \times 10^8$ m/sec (speed of light in vacuum).

Some of the features of *corpuscular excitation* of phosphors, such as by cathode rays, may be exemplified with the aid of Fig. 8, which shows generalized sketches of the interiors of phosphor crystals, including three major classes of crystal irregularities (faults). The total penetration, x_t, of 10^3 to 10^6 volt cathode rays in a phosphor of density σ (in grams per cubic centimeter) is calculable from Terrill's equation:

$$x_t = \frac{V_0^2}{4 \times 10^{11}\sigma} \quad \text{cm} \quad (V_0 \text{ in volts}) \tag{10}$$

The fraction W/W_0, representing the power dissipated up to distance x in the crystal, is given by Stinchfield's equation:

$$\frac{W}{W_0} = 1 - \gamma\left(1 - \frac{x}{x_t}\right)^{1/2} \varepsilon - \frac{C'x/x_t}{[1+(1-x/x_t)^{1/4}]^4} \tag{11}$$

where $\gamma \approx 1$ for $x/x_t < 0.5$ (γ increasingly exceeds unity for $x/x_t > 0.5$), and $C' \approx 32$. Over 50 per cent of the cathode-ray power is dissipated in the first quarter of the total penetration distance, and over 80 per cent is dissipated in the first half of the total penetration distance.

A few phosphors, such as ZnO:[Zn], may be excited by cathode rays with energies as low as 5 volts, but such low-voltage excitation is quite inefficient because the excitation energy is expended in the distorted surface layers of the phosphor crystals and the ratio of secondary to primary electrons is usually less than unity at such low primary voltages. Conventional, unmetallized cathode-ray-tube screens must have secondary-emission ratios equal to or greater than unity to maintain a positive potential with respect to the cathode. The efficient range of primary voltages is above about 1000 volts, and preferably above 10,000 volts for phosphors whose limiting potentials (voltage above which the secondary-emission ratio falls below unity) are above 10,000 volts or phosphor screens which may be coated with an electron-pervious reflecting and conducting coating, such as a 1000-Å-thick layer of aluminum.

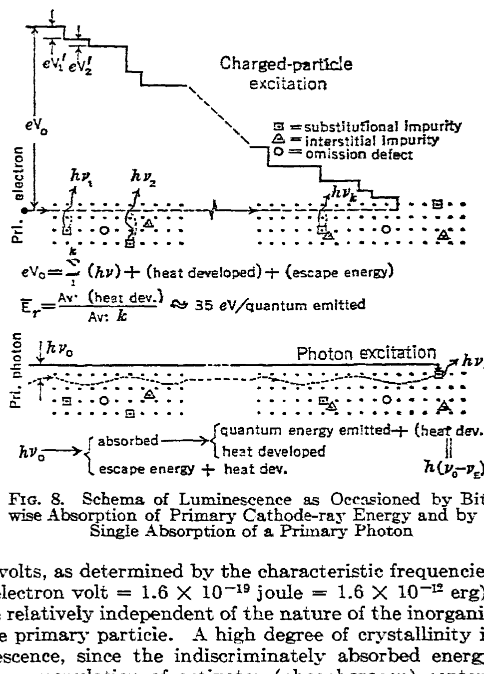


FIG. 8. Schema of Luminescence as Occasioned by Bitwise Absorption of Primary Cathode-ray Energy and by a Single Absorption of a Primary Photon

As indicated in Fig. 8, the initial energy, eV_0, of the primary particle is expended bitwise and indiscriminately to the crystal atoms and ions. The sizes of the absorbed energy bits average about 20 to 30 electron volts, as determined by the characteristic frequencies of bound electrons in the crystal (1 electron volt = 1.6×10^{-19} joule = 1.6×10^{-12} erg). The average absorbed energy bits are relatively independent of the nature of the inorganic material and the initial energy of the primary particle. A high degree of crystallinity is essential for efficient cathodoluminescence, since the indiscriminately absorbed energy bits must be transmitted to the sparse population of activator (phosphorogen) centers with the minimum of attenuation. Glassy structure, crystal faults, and undesirable impurities lower luminescence efficiency by converting absorbed primary energy into heat as indicated in eqs. (5) and (7). The efficiency loss (as heat) in eq. (8) is occasioned by the absorptivity of the phosphor crystal for its own luminescence. The more efficient cathodoluminescent materials require an average of over 30 electron volts of primary cathode-ray energy per 1.5 to 3 electron volt quantum of emitted luminescence. Hence, the efficiency of cathodoluminescent materials has thus far been less than about 10 per cent.

In *photon excitation* of phosphors, the primary photon, $h\nu_0$, seeks out a spot in the crystal to expend itself completely, not bitwise. The absorption of a primary beam, containing n_0 photons per unit cross-sectional area, is a function of penetration distance, x, into the phosphor, according to:

$$n = n_0 \varepsilon^{-Ax} \tag{12}$$

where the absorption coefficient A is strongly dependent on the frequency of the primary photon and the characteristic allowed frequencies in the phosphor crystal. The characteristic frequencies of bound electrons in the base materials comprising the bulk of phosphor crystals lie in the far ultraviolet ($\nu > 15 \times 10^{14}$ cycles per second, $\lambda < 2000$ Å)

Excitation of phosphors by near-ultraviolet photons results in direct excitation of the foreign activator centers, and there is no need for a high degree of crystallinity. Good photoluminescence, but inefficient cathodoluminescence, is obtained from amorphous materials such as organic dyes, inorganic glasses, and certain inorganic crystals such as the alkaline-earth-sulfide phosphors which contain some glassy structure caused by the residual non-volatile fluxes used in their preparation.

Photoluminescence efficiency is limited both by impurities and crystal faults, which convert primary photons and internal luminescence photons into heat according to eqs. (5), (7), and (8), and by the energy deficit, ΔE:

$$\Delta E = h(\nu_{\text{primary}} - \nu_{\text{emitted}}) \rightarrow \text{Heat} \tag{13}$$

Efficient phosphors, excited by near-ultraviolet photons, have photoluminescence efficiencies of the order of 60 to 80 per cent, with quantum efficiencies near unity. At very high energies of primary photons (or particles) the production of secondary radiations introduces added complications in the mechanisms of excitation of phosphors.

There are two major types of *luminescence-active centers* in phosphor cyrstals:

Substitutionally located activators are exemplified by manganese which replaces zinc, especially in phosphor crystals where oxygen or fluorine dominate the anion structure (e.g., zinc silicate, and zinc fluoride). These phosphors afford predominantly simple initial *exponential* (ε^{-at}) *decays* of light output, L, after excitation to a peak luminance, L_0, according to:

$$L = L_0 \varepsilon^{-at} \qquad (t \text{ in seconds}) \tag{14}$$

where the constant a has known values ranging from 10^6 for ZnO:[Zn] to 10 for $(Zn{:}Mg)F_2{:}Mn$. Exponential decays are determined largely by the chemical composition of the phosphor, being relatively unaffected by changes in crystal structure, in temperature, or in the type, duration, or intensity of excitation. The luminescence action is apparently localized in the substitutionally located activator sites as a metastable-state monomolecular process, without necessitating electron transport outside the immediate sphere of influence of the substitutional activator center. The observed weak photoconductivities of some ε^{-at}-decay phosphors are probably associated with their later-stage low-intensity power-law-decay "tails" which are strongly affected by changes in crystal structure, temperature, and excitation.

Interstitially located activators are exemplified by copper or silver interspersed among the regular lattice units in phosphors where sulfur and/or selenium dominate the anion structure. These phosphors afford so-called *power-law* (t^{-n}) *decays* represented by

$$L = L_0 \left(\frac{b}{b + t}\right)^n \tag{15}$$

where both b and n are not true constants but vary with changes in L_0, t, crystal structure, or temperature, and with the type and duration of excitation. For practical purposes, $0 < b < 10^{-3}$ and $0.2 < n < 3$ for known phosphors. There is a definite correspondence between the strong photoconductivities and phosphorescences of most t^{-n}-decay phosphors, indicating electron transport between the remote centers of trapping and emission. In these cases, monomolecular activated release followed by bi- or polymolecular mechanisms afford the complex t^{-n} decays.

The optimum activator concentrations and maximum allowable impurity concentrations in ε^{-at}-decay phosphors are about a hundredfold greater than those in t^{-n}-decay phosphors. Since the cube root of 100 is about 5, this means that interstitially located impurities are affected by other impurities five times as remote, as in the case of substitutionally located impurities. Substitutionally located impurities are in regions of lower potential energy and are better buffered by close coupling with the lattice forces.

The optimum number of active luminescence centers and/or electron traps is about 10^{17} per cubic centimeter in t^{-n}-decay phosphors and about 10^{19} per cubic centimeter in ε^{-at}-decay phosphors. From these data, and data on the penetrations and power densities of common excitants, it is possible to calculate the maximum phosphorescences of phosphors with given decays.

17. MECHANISMS OF SCOTOPHORS

Figure 9 shows a schematic section of a crystal of potassium chloride, KCl, indicating several omission defects and an interstitial defect. The fraction f of omission defects is a function of equilibrium temperature T (degrees Kelvin) according to

$$f = A\varepsilon^{-W_0/kT} \tag{16}$$

where, for KCl, $A \approx 10^4$ and $W_0 \approx 1.6 \times 10^{-19}$ joule. At temperatures near the melting point of KCl (1074 deg Kelvin), f becomes about 0.2, and many of the omission defects are frozen-in when KCl is evaporated and condensed, as in making P10 screens. Therefore, a large number of the 3.2×10^{22} lattice sites per cubic centimeter in P10-screen crystals are empty, and half of these omission defects are absent chlorine ions (Cl^-) whose vacant positions may be occupied by electrons to form scotophor color centers (so-called F-centers) as indicated in Fig. 9. The visible absorption band of tenebrescent KCl (Fig. 4) corresponds to absorption of light by F-centers, i.e., ejection of the trapped electrons. The maximum concentration of F-centers in KCl screens is about 10^{18} F-centers per cubic centimeter.

Tenebrescence is quantitatively expressed as contrast, C, defined by

$$C = \frac{100(L_0 - L_d)}{L_0} \quad \text{(in per cent)} \tag{17}$$

where L_0 and L_d are the luminances of the undarkened and darkened areas, respectively, when observed under light having wavelengths within the induced absorption band of the scotophor.

The decay of tenebrescence, i.e., the bleaching of induced darkening, is a power-law relation of the general type

$$C = C_0 t^{-n} \tag{18}$$

where C_0 is the tenebrescence (contrast) at time $t = 0$ and n decreases from about 1.5 to less than 0.1 with decreasing temperature or intensity of illumination and with increasing degree of tenebrescence. The useful decay intervals (time between successive excitations) of the best known scotophor, KCl, range between 1 and 60 seconds at temperatures near 40 deg cent and illuminations of the order of 10,000 foot-candles of light from incandescent lamps.

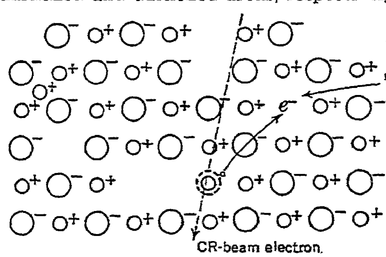


FIG. 9. Schematic Section of a Crystal of Potassium Chloride (KCl), showing the formation of an F-center. Large circles are Cl^-, small circles are K^+.

The sensitivity of cathodotenebrescent KCl is about $^1/_{100}$ that of efficient cathodoluminescent phosphors; i.e., perceptible tenebrescence requires about 100 times as much cathode-ray excitation energy as is required to produce perceptible luminescence.

18. SPECIFIC CHARACTERISTICS OF USEFUL PHOSPHORS AND SCOTOPHORS

Phosphors and scotophors may be considered as high-frequency transformers which can transform a megavolt-wide range of invisible photon or particle energies into emission or absorption of light in or near the 1.5-volt-wide visible region of the electromagnetic spectrum. The output frequencies of these materials are in the range from about 4×10^{14} to 8×10^{14} cycles per second (red to violet).

The chief commercial uses of phosphors are in electron discharge devices, such as fluorescent lamps and cathode-ray tubes for radar or television. *Fluorescent lamps* contain mercury vapor at a pressure of about 4 μ, which, under electron excitation, emits invisible 2537 Å radiation which, in turn, excites visible luminescence in an internal phosphor coating. White-emitting coatings for fluorescent lamps are made by mechanically mixing blue-emitting magnesium tungstate and yellow-emitting manganese-activated zinc-beryllium-silicate phosphors which afford quantum efficiencies of the order of 90 per cent and overall efficiencies four times as great as most incandescent lamps.

Cathode-ray-tube screens comprise a much greater range and variety of materials than those used in fluorescent lamps. Table 2 shows some *approximate* general properties of the more useful cathode-ray-tube screens, including all those presently coded by the Radio Manufacturers Association (RMA). The screens are arranged in the order of their approximate persistences after excitation by 6-kv cathode rays at about 1 microampere per square centimeter. The dark-trace screen at the right of Table 2 exhibits decelerated decay as the excitation density is increased; the sulfide-type bright-trace phosphor screens exhibit accelerated decays with increasing degree of excitation. By observing a tenebrescent image at some time after excitation, it is possible to obtain greater relative contrast between image traces than during the excitation interval. Conversely, the normal bright-trace screens may be used as contrast or ratio compressors.

Table 2. Approximate Data for Some Cathode-ray-tube Screens

Screens with Retentivities *Shorter* Than the Persistence of Vision

RMA No.	P15	P5	P11	P4, P6	P4	P1	P3
Material	$ZnO:[Zn]$	$CaWO_4:[W]$	α *-$ZnS:Ag$	$ZnS:Ag$ + $(Zn:Cd)S:Ag$	$ZnS:Ag$ + $Zn_8BeSi_5O_{19}:Mn$	α-$Zn_2SiO_4:Mn$	$Zn_8BeSi_5O_{19}:Mn$
Bright- or dark-trace	BT	BT	BT	BT	BT	BT	BT
Color	Green-white	Blue	Blue	White	White	Green	Yellow
Contrast	20:1	20:1	20:1	30:1	30:1	20:1	20:1
Candles/watt 6 kv, 1 $\mu a/cm^2$	1.8	0.2	2.1	4.5	2.4	3	2.5
Min. coulomb/cm^2 for visible trace	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}	10^{-10}
Operating (6 kv) millilamberts	12	1.3	14	30	16	20	17
Light source	CRT screen	CRT screen	CRT screen	CRT screen	CRT screen	CRT screen	CRT screen
Decay form	ϵ^{-at}(?)	ϵ^{-at}	t^{-n}	t^{-n}	t^{-n} $\downarrow$ 0.005 (B); ϵ^{-at} $\downarrow$ 0.06 (Y)	ϵ^{-at}	ϵ^{-at}
Decay time to 1% (sec) (approx.)	10^{-6}	10^{-5}	0.005	0.001 (Y) and 0.005 (B)	0.005 (B) 0.06 (Y)	0.05	0.06
Mechanism	Electronic	Electronic	Electronic	Electronic	Electronic	Electronic	Electronic
Range of scans/sec	$60-10^6$	$60-10^5$	$60-10^4$	$60-10^4$	$30-10^3$	$20-10^3$	$20-10^3$
Chief use	Photography of transients to 500 kc	Photography of transients to 60 kc	Photography of transients to 9 kc, 70 cm/μsec	Monochrome or color television. (P4 = bl. and wh., P6 = color.)	Projection television	Oscilloscopes and rapid-scan radars	Same as P1
Remarks	Efficient below 4 kv and above 5 $\mu a/cm^2$	Low efficiency	Highest total energy output	Se may be substituted in part for S. Yellow component gives over 6 cp/watt	Yellow component saturates less at high current densities	The "old reliable" for CRT	

Table 2. Approximate Data for Some Cathode-ray-tube Screens—Continued

"Memory" Screens; Retentivities *Longer* Than the Persistence of Vision

RMA No.	P13	P12	P2	P14	P7	P10
Material	$MgO:SiO_2:Mn$	$(Zn:Mg)F_2:Mn$	β *-ZnS:Cu(Ag)	β *-ZnS:Ag on ZnS(75)CdS:Cu	β *-ZnS:Ag on ZnS(86)CdS:Cu	KCl
Bright- or dark-trace	BT	BT	BT	BT	BT	DT
Color	Red	Orange	Green	White-orange	White-yellow	Magenta on white
Contrast	20:1	20:1	20:1	20:1	20:1	1.2:1
Candles/watt 6 kv, 1 $\mu a/cm^2$	0.3	2	4	2.1	2.4	....
Min. coulomb/cm^2 for visible trace	10^{-9}	10^{-10}	10^{-10}	10^{-9}	10^{-9}	10^{-7}
Operating (6 kv) millilamberts	2 (t = 0)	0.2 (0.2″)	0.1 (0.5″)	0.003 (2″)	0.03 (3″)	8000
Light source	CRT screen	CRT screen	CRT screen	CRT screen	CRT screen	External lamps
Decay form	$\epsilon^{-at} \rightarrow \epsilon^{-bt}$	ϵ^{-at}	t^{-n}	t^{-n}	t^{-n}	t^{-n}
Decay time to 1% (sec) (approx.)	0.1	0.4	0.3	1	3	5
Mechanism	Electronic	Electronic	Electronic	Electronic	Electronic	Electronic and ionic
Range of scans/sec	10-30	3-30	1-10	0.5-6	0.1-5	0.01-5
Chief use	Proposed for fire-control radars	Fire-control radars (4-16 scans/sec)	Long-persistence screens for oscilloscopes (prewar)	Eagle and H_2K radars ($\approx$ 1 scan/sec) Long-persistence oscilloscopes	Most radars operating slower than 1 scan/sec. (Especially PPI scans.)	Large projected images for slow-scan radars; less than 0.2 scan/sec
Remarks	Low efficiency, but stable relative to P12	Difficult to apply and unstable during operation	Largely superseded by P7 and P14	Two-layer cascade screen	Two-layer cascade screen	Bright image, but low contrast and unduly long decay

α * = cubic; β * = hexagonal. (B) = blue, (Y) = yellow.

Representative *spectral distribution curves* of the major cathode-ray-tube screens are shown in Fig. 10. The numbers near the peaks of the emission bands of individual phosphors are the heights of the bands relative to the peak of the cathodoluminescence band of α-Zn_2SiO_4:Mn (P1) which is arbitrarily set equal to 100. The visible cathodoluminescence efficiency of conventional P1 screens is about 3 candles per watt at 6 kv and 1 microampere per square centimeter. The spectral distribution curves of the white-emitting P4 (monochrome-television) and P6 (color-television) screens may vary considerably among different manufacturers, since there are many possible ways of producing equivalent-appearing white colors. The optimum white is yet to be decided by popular approval, although P4 screens are presently standardized at a color temperature of about 6500–7000 deg Kelvin. The P4 screen is usually a two-component mixture of blue-emitting and yellow-emitting phosphors [all-sulfide(selenide) or sulfide(selenide)-silicate]; the P6 screen is usually a three-component mixture of blue-, green-, and red-emitting sulfide- (selenide) phosphors.

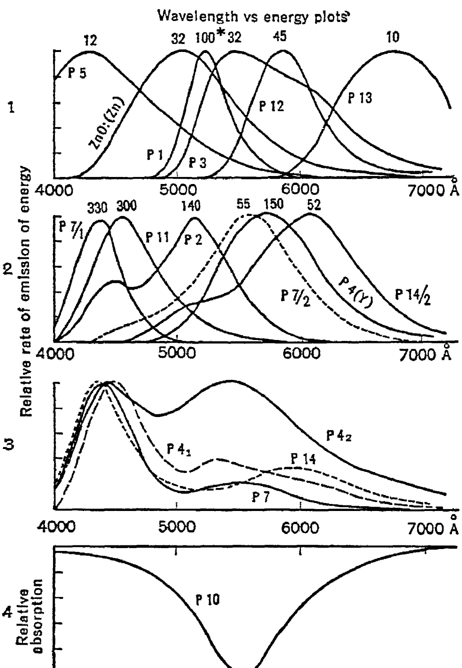


FIG. 10. Representative Spectra of Some Cathode-ray-tube Screens. P4 and P6 screens vary considerably among different manufacturers.

The optimum *screen thicknesses* for cathode-ray tubes vary from about 1 mg/cm² for phosphors having average particle sizes less than 1 micron, to 22 mg/cm² for cascade-screen (P7 and P14) phosphors having average particle sizes near 15 microns. The evaporated P10 (KCl) screen thickness is about 12 microns. These thicknesses are for operation below about 10 kv and must be increased for higher voltages according to eqs. (10) and (11).

The strontium-containing phosphors listed at the bottom of Table 1 are *infrared-stimulable materials* which have quite deep electron traps (of the order of 1 electron volt). These materials can store most of their latent phosphorescence energy for several days, even at room temperature. Efficient release of the stored phosphorescence energy is obtained by irradiating the previously excited material with near-infrared which may be varied in intensity to alter the normal concave-upward decay of the phosphor. These phosphors are good photoluminescent materials but are relatively inefficient under cathode rays. By cocrystallizing sulfate ions in copper-activated zinc sulfide phosphor, an unusual *infrared-quenchable phosphor* is produced which affords long phosphorescence plus the ability to decrease or terminate the phosphorescence by quenching with infrared at an arbitrary instant during the decay time. Infrared-sensitive phosphors may be used to store information for controllable time intervals or to convert positive images into negative images (or vice versa).

The intensity and duration of phosphorescence after cathode-ray excitation is generally less than that obtained after excitation by photons. Cathode-ray-excited phosphorescence may be increased by using the cascade principle. The cascade principle involves constructing a stratified-layer screen wherein an efficient photophosphorescent material is covered with a layer of an efficient cathodoluminescent material whose emission band overlaps the excitation band of the photophosphorescent material. By this method, cathode-ray energy is intermediately converted into photon energy which is more efficient in exciting phosphors to produce phosphorescence. Cascading may be used also to excite cathode-ray-modulated photoluminescence in materials, such as the infrared-stimulable phosphors, which are inefficient under direct cathode-ray excitation. The P7 and P14 screens are examples of practical cascade screens which were devised for radar cathode-ray tubes.

BIBLIOGRAPHY

1. Leverenz, H. W., Cathodoluminescence as Applied in Television, *R.C.A. Rev.*, Vol. V, No. 2, 131–176 (October 1940).
2. Leverenz, H. W., Luminescence and Tenebrescence as Applied in Radar, *R.C.A. Rev.*, Vol. VII, No. 2, 199–240 (June 1946).
3. Some of the data mentioned herein were obtained during the course of Contracts NDCrc-150 and OEMsr-440 between the Office of Scientific Research and Development and the Radio Corporation of America. A more complete account, with over 400 further references, may be found in H. W. Leverenz, Final Report on Research and Development Leading to New and Improved Radar Indicators, *Report NDRC-14-498*, June 30, 1945, obtainable from the Office of the Publication Board, Department of Commerce, Washington, D. C. (PB25481).
4. Leverenz, H. W., Luminescent Solids (Phosphors), *Science*, Vol. 109, No. 2826, 183–195 (Feb. 25, 1949).
5. Leverenz, H. W., *An Introduction to Luminescence of Solids*, Wiley (1950).
6. Fonda, G. R., and Seitz, F. (Editors), *Preparation and Characteristics of Solid Luminescent Materials*, Wiley (1948).

CATHODE-RAY TUBES

By L. E. Swedlund

The cathode-ray tubes considered herein are those electron tubes in which a relatively low-current electron beam of small circular cross-section is focused on, and deflected across, a luminescent screen. Their principal applications are in television receivers, oscillographs, and radar-type indicators. Their great value is the ability to display variations in voltage and current without the limitations of mechanical inertia, and their relatively low cost. Except for a few specialized designs, they are sealed-off, high-vacuum tubes, having an electron gun with a heater-type, oxide-coated cathode located in the neck of a cone-shaped glass bulb. A luminescent screen is deposited in the large, nearly flat end of this cone. Figure 1 shows an electrostatic-focus, electrostatic-deflection oscillograph-type cathode-ray tube, and Fig. 2 a magnetic-focus, magnetic-deflection tube, for television image reproduction. Magnetic focus is very seldom used with electrostatic deflection, but electrostatic focus is often used with magnetic deflection. The principal parts of the cathode-ray tube are the electron gun, the bulb, and the luminescent screen.

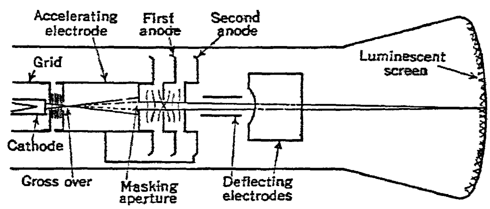


FIG. 1. Electrostatic-focus Electrostatic-deflection Cathode-ray Tube

19. ELECTRON GUN

The design of the electron gun is based on the principles of electron optics outlined in Section 14. However, owing to the complex nature of the electron paths, particularly in the region of the grid and cathode, and to the large number of interrelated factors, much of the design is based on experimental information. The active cathode surface is a small, flat, oxide-coated nickel surface normal to the gun axis. Directly heated, high-temperature cathodes are occasionally used in high-voltage tubes to withstand better the effects of positive-ion bombardment. A control grid having a round aperture of approximately 0.04-in. diameter is spaced as closely as practical to the cathode. As the spacing may amount to only 0.002 in. when the cathode is hot, it is difficult to hold this spacing constant from tube to tube. This small variation in spacing results in a fairly large variation in control-grid voltage for beam-current cutoff. The control grid is nearly always designed to operate at a negative potential with respect to the cathode. The beam is accelerated by means of a screen grid, first anode, and second anode; by a screen grid and anode; or directly by an anode to full-beam potential. In each case, the

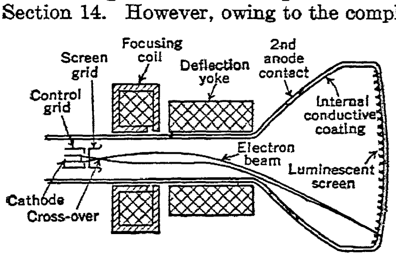


FIG. 2. Magnetic-focus Magnetic-deflection Cathode-ray Tube

spacing between the first accelerating electrode and the control grid is adjusted to provide about the same accelerating field at the cathode. Maximum beam currents of the order of 1 milliampere are drawn at zero bias, and the specified control-grid bias for beam cutoff varies from approximately 20 to 100 volts.

OPERATION OF THE ELECTRON GUN. The field above the cathode draws the electrons into a focus in a few millimeters' travel. In this field they cross over, and beyond it they travel in straight paths until they enter the final focusing field near the end of the gun. Here the electrons are made to converge to a second focus or cross-over at the screen. The final focusing field or electron lens may be either electrostatic or magnetic. The electrostatic lens consists of coaxial cylinders, apertured disks, or a combination of both. These components are always mounted inside the tube. The magnetic lens is always mounted outside the neck of the tube and usually is a coaxial coil encased in an iron shell except for a short axial magnetic gap on the inside surface. In both cases a very good degree of rotational symmetry is required. The usual error due to any lack of symmetry is astigmatism, which results in an elliptical shape when the spot is focused. Electrostatic focus is adjusted by varying the first-anode voltage. The first-anode voltage is usually about one-fifth the second-anode voltage. Magnetic focus is adjusted by changing the magnetic field in the region of the electron beam. This is done by varying the current in an electromagnetic focus coil or by varying a magnetic shunt in a permanent-magnet focus unit.

ELECTROSTATIC DEFLECTION. Figure 1 shows two pairs of deflecting plates at right angles to each other at the exit of the electron gun. The average potential of each pair should be near that of the second anode of the gun since a difference will produce a focusing effect. Since such focusing produces a strong astigmatic effect it is sometimes possible to adjust their average potential to counteract astigmatism in the electron gun. Most tubes, however, are so well made that such correction is not needed. The best uniformity of focus is attained with symmetrical or push-pull deflection. It is possible to connect one plate of a pair to the second anode and apply voltage only to the other at the expense of focus and sometimes linearity of deflection. When the deflection plates are parallel the deflection can be calculated from the following equation.

$$h = \frac{LlV_d}{2V_bd} \qquad (1)$$

where h is the deflection from the center, L is the plate to screen distance, l the deflecting-plate length, V_d the difference in potential between the plates, V_b the second-anode or beam voltage, and d the spacing between plates. In order to provide maximum sensitivity, the plates are usually shaped to follow the contour of the deflected beam. The expression for deflection becomes more complex but has about the same proportionality factors.

MAGNETIC DEFLECTION. Magnetic deflection is accomplished by placing a magnetic field at right angles to the beam path. This field is usually provided by a pair of coaxial coils placed on opposite sides of the neck in order to produce a nearly uniform deflecting field. The beam moves at right angles to the magnetic field in a direction indicated by the familiar left-hand motor rule, remembering, however, that conventional current flow is opposite that of electrons. Unlike electrostatic deflection, it is possible and advantageous to place a second pair of coils for orthogonal deflection in the same region as the first. A combination of deflection coils is known as a deflection yoke and is ordinarily designed to meet the rather specialized requirements of television. Magnetic pole pieces and return paths are sometimes provided to decrease the current required by reducing the reluctance of the magnetic path. The deflection h at the screen for a magnetic field of length l having uniform flux density of B at a distance L from the screen, and for a final voltage E_b, is

$$h = \frac{0.298 \times 10^{-6}BlL}{\sqrt{V_b}} \qquad (2)$$

Since there is a large air gap in these coils, the flux density is proportional to the current and thus deflection is proportional to the current. Although some approximations have been made, this equation holds well up to 60° total deflecting angle. It is seldom desired that the field be uniform across the neck because the screen is usually not properly curved to produce a pattern which looks rectangular. In the reflective optics projection system, for example, the pattern is made to "barrel" slightly to correct for "pin-cushion" distortion in the optics (see Section 14).

COMPARISON OF MAGNETIC AND ELECTROSTATIC DEFLECTION. In comparing eqs. (1) and (2) for electrostatic and magnetic deflection, it is to be noted that, with an increase of beam voltage, deflection decreases with the anode voltage in the former and as the square root of the anode voltage in the latter. Thus, magnetic deflection becomes relatively more favorable as the voltage is raised. In television, this, plus relatively better focus and economic factors, sets the dividing point at about 5 kv. Electrostatic and magnetic deflection are occasionally combined in one tube, though tubes designed specifically for this type of operation have not found many applications.

DEFLECTION DEFOCUSING. Deflection is accompanied by a deterioration of focus known as deflection defocusing. It appears as an elongation of the spot in the direction of deflection in electrostatic deflection and as an elongated and rotated spot in magnetic deflection. It increases approximately as the square of the deflecting angle. Thus deflection distortion limits the deflecting angle and indirectly the beam current because the amount of current which can be focused into a given size spot is determined by the size of the beam as it leaves the gun. For high currents and wide deflecting angles the deflection distortion is five to ten times as great with electrostatic deflection as with magnetic deflection. As a result, electrostatic-deflection tubes are designed for smaller focused currents and narrower deflecting angles than magnetic-deflection tubes. The change of focus due to the increase of gun to screen distance with deflection is small compared to deflection defocusing even with flat screens. But, since deflection defocusing also produces a shortening of the image focal distance, it is possible to make a small amount of compensation by slightly underfocusing the beam at the center of the screen. This effect is usually so small that it is not worth while to modulate the focus voltage in synchronism with scanning to improve the uniformity of focus.

20. BULBS FOR CATHODE-RAY TUBES

The bulb of the cathode-ray tube must meet specialized requirements for mechanical strength, dimensions, optical quality, and electrical insulation. Special grades of glass are needed to meet these demands. Although the bulb has to withstand only atmospheric pressure (15 lb per sq in.) it is good practice to design for a breaking strength of 60 lb or more per sq in. This is done to avoid failure due to slight scratches, temperature shock, and aging. Great care should be exercised, particularly in handling the larger tubes, to protect personnel from flying glass in the event of accidental implosion. Generally the zone of highest stress in a bulb is near the rim of the face. Here the outside surface of the glass is in tension, in which it is weak, and so particular care should be taken to avoid scratches and shocks in this zone. In the past large bulbs were usually made of heat-resisting, low-thermal-expansion glasses, but because of both cost and optical quality most bulbs are now made of a soft glass. For all but the low-voltage oscillograph-type tubes, this is a lead-type glass in contrast to the usual lime-type soft glass. It is chosen to obtain high electrical resistivity. Large bulbs are mass-produced by pressing parts which are fused together. If the surface must be very good and accurately shaped, as in projection type tubes, the face plate may be ground and polished before sealing to the bulb.

LIMITATIONS IMPOSED BY THE LUMINESCENT SCREEN. The luminescent screen is a very important part of the cathode-ray tube. (See articles 15–18.) The limitations and requirements of the luminescent screen have an important bearing on gun-design problems. Both the amount of current that can be sharply focused and the amount that can be utilized efficiently by the screen are limited at low voltage. High brightness, consequently, has to be obtained by raising the beam voltage applied to the screen. Although raising the anode voltage improves the focus and screen efficiency, it also increases electrical insulation problems, introduces screen charging, and requires higher deflection energy. Specially developed, low-current, high-voltage power supplies which have low cost, compactness, and safety have made this problem less difficult.

SCREEN SIZE AND BRIGHTNESS. When the screen size is increased without changing the gun or deflecting angle, the anode voltage must be raised to maintain the same number of lines resolution and brightness. This is due to the fact that the size of the focused spot increases with bulb length and that the beam energy per unit of screen area determines the brightness. Projection tubes require the highest screen brightness and are therefore operated at the highest voltages. Luminescent screens are very thin and thus have a very small thermal storage capacity. Great care must be taken not to move the beam slowly or over too small an area. With a small overload the screen may be damaged only temporarily, as by overheating, or to a light "burn" which gradually fades out. The total time of screen bombardment during the life of a tube is surprisingly small. Consider a television tube scanned with a 525-line raster. This provides roughly 360,000 picture elements; and so, neglecting blanked-out return line time and modulation, each element of the picture area is being bombarded only 1/360,000th of the time. One thousand hours amounts to 3,600,000 seconds. Thus in a representative tube lifetime, the screen has been bombarded less than 10 seconds. However, during this short time it is bombarded at a very high momentary energy density; for example, in a 5-in. projection tube it amounts to 25 kw/cm^2.

DISCHARGE OF SCREEN. Luminescent screen materials are good electrical insulators and are prevented from charging up negatively under electron bombardment only by virtue of a secondary-to-primary electron-emission ratio equal to or greater than unity. The range of voltage where this is true varies with the type of screen material, its manufacture, and application. Under favorable conditions screens can be discharged by secondary emission over the 500 to 25,000 volt range. At the upper end of the range not all phosphors can be used and a charge of several thousand volts may build up on the screen, thus reducing the effective screen voltage. In general, secondary emission becomes poorer at low current density and with use. Poor secondary emission at low voltage causes the screen to charge up enough to deflect the beam off the screen. For example, trouble is likely to occur if beam current flows to the screen while the anode voltage is building up. A charge is then built up before the screen is bombarded with full-voltage electrons. At the upper end of the range the light output fails to increase at the normal rate with voltage increase and the pattern may become unstable. The effects of poor secondary emission may be overcome by mounting the phosphor on a conducting surface and, above about 7000 volts, by placing a very thin, light metal back on the luminescent screen. The metal back is especially desirable because it can be made light-reflecting in order to utilize the light ordinarily lost through the back of the screen. It can also serve as an effective barrier to negative ions.

ION SPOT IN SCREEN. Ion spot in the screen is a problem, particularly in magnetically deflected tubes. It is the area in the luminescent screen which has been damaged by negative-ion bombardment. The screen is damaged almost altogether by the negative ions which form in or near the cathode surface when it is bombarded by positive ions. The rate of damage increases with voltage, but it is most noticeable at about 7 kv because above this voltage the electrons begin to penetrate below the damaged surface layer of the screen. It is greatly affected by the amount and kind of gas left in the tube, but it is almost impossible to eliminate ion damage completely by evacuation and gas getters. Because electrostatic focusing is independent of mass and magnetic focusing is not, electrostatic focus guns focus the ions, thus forming small ion spots rather quickly whereas magnetic focus tubes form large ion spots slowly. Initially the small spot is more objectionable, but with time the large spot becomes more objectionable. In addition to the metal-backed screen, which is rather expensive to apply to low-voltage tubes, there is another effective solution known as the "ion trap" electron gun. Figure 3 shows two constructions of this idea. Both are simplified modifications of the mass spectrograph. The combined beam of ions and electrons is directed off the center of the gun into a trap in *A* by bending the anode and accelerator and in *B* by means of a tilted electrostatic lens. In both, a magnetic field is applied to restore the electrons to the axis. The secondary magnetic field used in construction *B* is useful not only in aligning the beam with the gun axis but also in overcoming the slight mechanical misalignments that may be present.

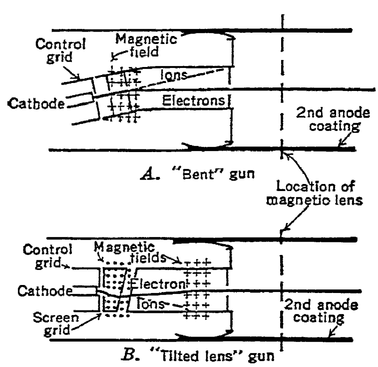


FIG. 3. Sectional Views of Ion Trap Guns

21. CHARACTERISTICS OF THE IMAGE

SPOT SIZE. The image detail or resolution is determined principally by the size of the focused electron image on the screen. Contrast and grain size of the screen may also be factors. The spot does not have a definite edge but decreases in brightness from the center approximately as a segment of a sine wave. When the beam scans a television-type raster it is found that, when the lines overlap to the point where the brightness is half that of the center, the line structure practically disappears. For this reason it is convenient to express spot size as the width at these half-brightness values. This is the basis of the "compressed raster" method of measuring spot size sometimes used for oscillograph-type tubes. However, the rate at which the brightness falls off toward the edge of the spot is important in determining the resolving power, and so for television reproduction tubes a resolution pattern signal which takes this into account is generally used. (See Fig. 10, p. 15-25.)

The resolution pattern has sets of spaced black and white lines usually converging in wedges along with circles and bars for other tests. The limiting resolution is defined as the point where the black and white lines merge. It is important that the black to white grid signal level be maintained beyond the value of resolution being read. If the reproduction tube is not to limit the resolution of the television signal noticeably, its limiting resolution should be about three times that of the television signal. The resolution is best at the center of the raster and poorest in the corners, and since it depends on the deflection yoke it may vary with the direction of scanning. Its value is usually expressed in terms of the number of black and white lines required to fill the vertical height of the raster. In photographic and optical literature it is generally the practice to count only the number of black lines per unit of length.

CONTRAST. The ability to see an image on a screen depends not only on the brightness of the trace but also on the ratio of this brightness to the background brightness. This is known as the brightness contrast ratio. This ratio, also sometimes referred to as contrast range, determines the number of distinct half tones which can be reproduced. Its value depends not only on the inherent characteristic of the screen but also upon the area and distribution of illumination in the screen. The minimum or limiting value of contrast ratio can be found by uniformly scanning all but a small black spot in the center of the raster. A contrast ratio representative of large areas can be found by scanning half the picture area and measuring the brightness in the center of each half.

GAMMA. The term gamma is borrowed from photographic practice to designate the half-tone characteristics of a television system. Over the brightness range useful for television reproduction the eye recognizes brightness differences as percentage increases in steps of 1 to 2 per cent. For this reason it is convenient to plot the grid drive versus light output of a picture tube on log-log scales. The slope of this curve is known as the gamma. In the same way each part of the system from the light in the scene on through the amplifiers has a gamma value. The overall gamma is the product of those of the series of elements. Unity overall gamma, which is usually desired, then represents brightness values in the image proportional to those in the scene. The gamma should be uniform over the whole brightness range for accurate reproduction, but it is usually less at the low and high brightness levels, as is also true in photographic reproduction. It is the practice in photography to refer to gamma as "contrast." This should not be confused with contrast ratio.

It has become the practice to mark the d-c bias and the signal controls of television receivers as "brightness" and "contrast" respectively. However, both controls affect the gamma and brightness. Overbiasing increases gamma by eliminating detail in the shadows; underbiasing decreases it by "washing out" shadow detail. Average brightness varies with d-c bias. Increase in brightness by signal increase is usually limited by the increase in spot size or amplifier overload. This loss of detail and contrast in the highlights is sometimes referred to as "blooming." In a picture tube the average gamma varies from 3 to 2, depending on the amount of masking in the gun. Correction to match the gamma of a standard transmitted signal must be done by introducing suitable curvature in the amplifier input versus output characteristic.

POST ACCELERATION. The beam is sometimes accelerated after deflection in order to increase the spot brightness without a proportionate increase in deflection energy. For this purpose the bulb conductive coating is separated and brought out to a third anode terminal. It has been applied only to electrostatic deflection tubes because little or no gain results when it is used with magnetic deflection. Post acceleration results in a focusing action which tends to bend the deflected beam back to the axis. There is also a small amount of focusing action on the spot which is readily corrected by raising the first anode voltage a small amount. In a typical post-accelerator-type oscillograph tube the deflection voltage must be raised 20 per cent when the third anode is increased from equal to double the second anode voltage. This increase in deflection voltage is accompanied by an increase in deflection defocusing so that the sharpness of focus is not as good as an equivalent non-accelerator tube whose screen voltage is the same. The deflection sensitivity also will not be as much less as indicated on the comparative tests on the accelerator tube because the deflection plates can be designed for closer spacing. However, the gain in brightness and deflection sensitivity along with the fact that the high-voltage power supply is grounded between its negative and positive terminals often justifies the higher cost of a tube of this type. The accelerator ratio is ordinarily limited to about 2, but where very high spot brightness is needed tubes may be designed to operate up to a ratio of 10 times.

PHOTOGRAPHY OF CATHODE-RAY-TUBE TRACES. A permanent record is sometimes needed, and when a tracing of the pattern on the tube cannot be readily made it can usually be photographed. Unless a time exposure can be made the brightness level suitable for viewing must usually be increased by raising the beam current and possibly by raising the screen voltage. Luminescent screens such as the P5 and P11 are particularly suited to photography. The blue light of these screens provides energy where the film is most sensitive. The lower-cost "color blind" films may sometimes serve nearly as well as the orthochromatic and panchromatic films. Since the P11 screen has approximately 10 times the photographic and visual efficiency of the P5 screen it is preferred, but occasionally its longer persistence may be a limitation. The proper exposure depends on lens speed, film sensitivity, screen size, spot brightness, and spot velocity. Tables suitable for estimating the best exposure are available from tube manufacturers, but usually a suitable value can be found from a few test exposures. Single television frames can be photographed with a hand camera having a fast lens and by using high-speed panchromatic film. For motion-picture records the camera must be exposed in synchronism with the frame frequency.

22. TELEVISION CATHODE-RAY REPRODUCTION TUBES

Cathode-ray tubes for television image reproduction are known as picture tubes or kinescopes. In directly viewed tubes the size of the screen largely determines the cost and operating voltage. Below a diameter of 7 in. the reduction in cost is not great and the image size is too small for comfortable viewing. Ten-, fifteen-, and twenty-inch bulbs provide images in steps of double-sized areas. Intermediate sizes such as 12 in. are also used. In the largest bulbs the economic limit for evacuated glass bulbs is reached. Larger images have to be produced by optical projection from a small screen tube. This small tube, known as a projection tube, must produce a much brighter image to illuminate the larger viewing screen and to make up for the losses in the optical system and so is operated at a higher anode voltage.

Table 1. Television Picture-reproduction Cathode-ray Tubes

Type	Size, in.		Typical Operation								Comments
	Nominal Diameter	Length	Heater Volts	Heater Amperes	Anode 2, kv	Anode 1, volts	Grid 2, volts	Control Grid Cutoff	Focus	
5TP4....	5	11 3/4	6.3	0.6	27	4900	200	−70	ES	Reflective optics projection
7CP4....	7	13 7/16			7	1365	250	−45	ES	Monitor use in daylight
7DP4....	7	14 1/16	6.3	0.6	6	1430	250	−45	ES	Small size receiver
7EP4....	7	15 1/2	6.3	0.6	2.5	650		−60	ES	ES deflection, low-cost receiver; see oscillograph tubes for deflection factors
7JP4....	7	14 1/2	6.3	0.6	6	1000		−120	ES	ES deflection for low-cost receiver; see oscillograph tubes for deflection factors
7HP4....	7	13	6.3	0.6	6		250	−55	M	Small size receiver
10AP4...	10	17 1/2	6.3	0.6	6			−60	M	Bent gun ion trap
10BP4...	10	17 5/8	6.3	0.6	9		250	−45	M	Directly viewed, tilted lens ion trap
10CP4...	10	16 5/8	6.3	0.6	8		250	−45	M	Directly viewed
10EP4...	10	17 5/8	6.3	0.6	8		250	−45	M	Directly viewed, tilted lens ion trap
10FP4...	10	17	6.3	0.6	9		250	−45	M	Directly viewed, metal-backed screen
12JP4...	12	17 1/2	6.3	0.6	10		250	−45	M	Directly viewed
15AP4...	15	20 3/16	6.3	0.6	10		250	−45	M	Large directly viewed
15BP4...	15	19	6.3	0.6	12		250	−45	M	Large directly viewed, rectangular face
20BP4...	20	28 3/4	6.3	0.6	15		250	−45	M	Large directly viewed

Note: The above table includes the tubes commercially available. Developmental tubes and those recommended for replacement only are not included. However, some of the tubes in these lists are in fact outmoded and will in time become obsolete. Owing to changes and additions which are constantly being made, up-to-date and more detailed information should be obtained from tube manufacturers or their agents.

The deflecting angle in all the above tubes is 50° or slightly above. All use P4-type screens.

DEFLECTION AND FOCUS OF PICTURE TUBES. Magnetic deflection is used for all but the smallest television picture tubes because it permits the use of wider deflection angles and higher beam currents. Owing to the desire for short overall length, particularly in the large tubes, the deflecting angle is made as large as practicable. Fifty-degree total angle has been found to provide a good compromise between focus uniformity, beam current, and deflecting power. Both magnetic and electrostatic focus guns are used, the former being favored for directly viewed tubes and the latter for projection-type tubes. At the high anode voltages used in the projection tube, energy for the focusing coil becomes a more important factor and the coil is also difficult to mount in the reflective optics system generally used for projection. The highest-voltage power supplies tend to have the poorest voltage regulation, so that a first anode which can be held in focus by means of a tap on the bleeder across the high-voltage supply is quite desirable. The bulb size of projection tubes is limited by the cost and bulk of the optical system used with them. For home use a 5-in. bulb and 27 kv appears to be an economic choice. For intermediate and theater-size images larger tubes and much higher operating voltages are justified.

23. OSCILLOGRAPH-TYPE CATHODE-RAY TUBES

These tubes are nearly always of the electrostatic-focus, electrostatic-deflection type. Their largest field of application is in low-voltage oscillographs for visual observations. These operate in the range of 500 to 2000 volts, but, for the observation and photography of high-frequency transients, anode voltages up to 30 kv may be used. The higher brightness and sharper focus afforded by higher-voltage operation are desirable but it results in lower deflection sensitivity and higher cost tubes and accessories. The signal must usually be amplified, and because a wide band width is desired the gain per stage is low. The size of the screen also influences the choice of anode voltage, since the energy to the screen should go up with its size. Conversely, with a given anode voltage, a small-diameter tube may indicate more detail than a large screen tube. Of the factors determining oscillograph-tube performance, deflection sensitivity is the most important. Deflection sensitivity is customarily expressed as the number of millimeters' deflection at the screen per volt (d-c) of deflection voltage. The reciprocal of this term, the deflection factor, is usually expressed in d-c volts per inch of deflection. The latter term is favored for design work. Spot size and contrast are next in importance. Grid modulation sensitivity is of relatively less importance than in television. It is desirable that the focus remain sharp while the beam current is varied, so most oscillograph-type tubes utilize an electron gun requiring essentially zero-first-anode current in order to avoid the effects of voltage regulation in the first anode circuit.

HIGH-FREQUENCY DEFLECTION. When the deflecting electrode leads must have minimum inductance, capacitance, and coupling to other leads, they are brought directly through the neck of the bulb. This is also of value when very wide frequency band amplifiers are used. For frequencies of about 100 megacycles and above it is necessary to take into account the electron transit time through the plates. Special tubes with short-length plates and high beam voltage have been used to record frequencies as high as 10,000 megacycles per second.

MULTIPLE GUN OSCILLOGRAPH TUBES. Oscillograph-type tubes with two or three independent guns and deflection systems are available for simultaneously producing two or more traces. Owing to their more complex construction their cost is higher than that of the equivalent number of single tubes. An electronic switch or the simultaneous use of two or more tubes is sometimes satisfactory.

RADIAL DEFLECTION TUBES. Radial deflection tubes have a rod-type deflecting electrode extending through the face into the middle of the cone. The deflection sensitivity of the rod decreases as the square of the deflection so that it cannot readily be calibrated. It is used principally to make a marker pip on a circular time base.

BIBLIOGRAPHY

Law, R. R., Contrast in Kinescopes, *Proc. I.R.E.*, August 1939, pp. 511–524.
Bachman, C. H., Image Contrast in Television, *Gen. Elec. Rev.*, Vol. 48, No. 9, 13–19 (September 1945).
Feldt, Rudolf, Photographing Patterns on Cathode-ray Tubes, *Electronics*, February 1944.

Table 2. Cathode-ray Tubes and Their Characteristics

Oscillograph-type Tubes

Electrostatic-deflection Types

Type	Size, in. Nominal Diameter	Size, in. Length	Typical Operation: Heater Volts	Heater Amperes	Anode 3	Anode 2	Anode 1	Control Grid Cutoff	Deflection Factor d-c volts/in. $D_1 - D_2$	Deflection Factor $D_3 - D_4$	Other Screens Listed	Comments
913(P1)	1	4 3/8	6.3	0.6		400	80	−32	240	176		Small indicator
902, 902A	2	7 7/16	6.3	0.6		600	150	−60	139	117		Oscillograph (replacement)
2AP1, 2AP1A	2	7 7/16	6.3	0.6		1000	250	−60	250	196		Oscillograph
2BP1	2	7 5/8	6.3	0.6		1000	215	−48	135	87	11	Oscillograph
3AP1, 3AP1A	3	11 1/2	2.5	2.1		1500	430	−50	114	109	All	Oscillograph
3BP1, 3BP1A	3	10	6.3	0.6		2000	575	−60	200	148	All	High-altitude operation
3DP1, 3DP1A	3	10 7/16	6.3	0.6		1500	430	−45	150	111		Radial deflection; altimeter
3EP1	3	10 3/16	6.3	0.6		2000	575	−60	221	165	2, 4, 5, 7	Oscillograph
3FP7, 3FP7A	3	10	6.3	0.6	4000	2000	575	−60	250	180	1, 2, 4, 5	Aircraft radar indicator
3GP1	3	11 1/2	6.3	0.6		1500	350	−50	120	105	4, 11	Oscillograph
3JP1	3	10	6.3	0.6	4000	2000	575	−60	200	148	2, 4, 7, 11	High-intensity oscillograph, aircraft radar
3KP1	3	11 1/2	6.3	0.6		1500	345	−48	89	60		Oscillograph
908A (P5)	3	11 1/2	2.5	2.1		1500	430	−50	114	109		Short time lag indicator
5AP1	5	13 3/8	6.3	0.6		2000	575	−57	93	90	2, 4, 5, 7	Short oscillograph
5BP1, 5BP1A	5	16 3/4	6.3	0.6		1500	337	−30	63	57	All	Oscillograph
5CP1, 5CP1A	5	16 3/4	6.3	0.6	4000	2000	575	−60	92	79	All	High-intensity oscillograph, aircraft radar
5JP1, 5JP1A	5	16 3/4	6.3	0.6	4000	2000	520	−75	96	96	All	H-f oscillograph and radar deflection leads through neck
5LP1	5	16 3/4	6.3	0.6	4000	2000	500	−60	103	90	All	Oscillograph
5RP1	5	16 3/4	6.3	0.6	20,000	2000	575	−60	175	164	2, 4, 7, 11	High writing rate photography and observation; projection oscillograph
5SP1	5	18 1/4	6.3	0.6	4000	2000	575	−60	92	79	2, 4, 7, 11	Two independent and identical guns; oscillograph
5UP1	5	14 3/4	6.3	0.6		1500	360	−48	51	41	7, 11	Oscillograph
905A(P1)	5	16 1/2	2.5	2.1		1500	338	−26	86	83		Oscillograph, deflection leads through neck
912(P1)	5	16 1/2	2.5	2.1		10,000	2000	−60	645	500		High-intensity oscillograph leads through neck
7EP4	7	15 1/2	6.3	0.6		2000	520	−48	88	76	1	Oscillograph, television, short length
7JP4	7	14 1/2	6.3	0.6		2000	667	−40	72	59	1	Oscillograph, television, short length
914A(P1)	9	20 1/16	2.5	2.1		2500	515	−50	115	91		Oscillograph, deflection leads through neck
12GP7	12	22	6.3	0.6	6000	3000	857	−98	89	83		Oscillograph and radar

Note: RMA tube numbers followed by "A" indicate improved designs which are interchangeable with the old; however, the reverse is not necessarily true. In the case of the electrostatic-deflection tubes it indicates the change to zero-first-anode current gun.

Magnetic-deflection-type Oscillograph Tubes

Type	Size, in. Nominal Diameter	Size, in. Length	Typical Operation: Heater Volts	Heater Amperes	Anode 2	Anode 1	Grid 2	Control Grid Cutoff	Focus	Deflection Angle	Comments
5FP7, 5FP7A	5	11 1/8	6.3	0.6	4000		250	−50	M	55°	Radar indicator
7BP7, 7BP7A	7	13 1/4	6.3	0.6	4000		250	−50	M	55°	Radar indicator
7CP1	7	13 7/16	6.3	0.6	4000	1365	250	−45	ES	55°	Television monitor
9GP7	9	17	6.3	0.6	7000		250	−50	M	55°	Radar indicator
9LP7	9	14 31/32	6.3	0.6	7000		250	−80	M	55°	Radar indicator
12DP7, 12DP7A	12	20 1/8	6.3	0.6	7000		250	−50	M	55°	Radar indicator

24. CATHODE-RAY-TUBE DISPLAYS

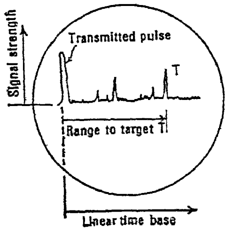
By T. Soller

A cathode-ray-tube display is a means of presenting information concerning a signal (which may be any quantity that can be put into the form of an electrical voltage) in terms of one or more independent variables on which the signal depends. Usually at least one of these variables is an explicit or implicit function of time. The variables are represented by displacements along cartesian or polar coordinate axes; the displacements are usually periodic, a single traversal along any axis being called a "sweep."

A display is said to be deflection-modulated if the signal produces a lateral displacement of the electron beam from some base line or sweep; it is said to be intensity-modulated if the signal voltage is used to increase or decrease the intensity of the electron beam. A deflection-modulated display is used when quantitative information regarding the nature of the signal is required, for an intensity-modulated display can at best yield only qualitative information regarding the signal strength or shape. An intensity-modulated display offers the advantage of being able to present the signal in terms of two variables, rather than the single variable (usually time) of the deflection-modulated display. Well-known examples of these two general types of display are the usual form of cathode-ray oscillograph (deflection-modulated) and the television display (intensity-modulated).

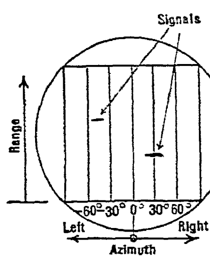
Many different types of display were developed in connection with the applications of radar during World War II. In many of these an attempt was made to present on the two-dimensional surface of the CRT screen information regarding the signal involving three variables, for example, range, azimuth, and elevation. The designation and geometry of these various display types are shown in the following pages.*

* For a more complete discussion of the subject, see: M.I.T. Radiation Laboratory Series, Vol. 22, *Cathode Ray Tube Displays*, McGraw-Hill Book Co., New York, 1947.



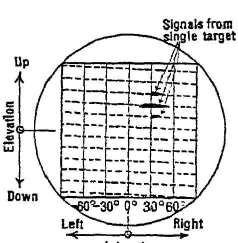
Type A

A deflection-modulated display, commonly used with electrostatic-deflection types of cathode-ray tubes. Consists of a linear horizontal time base, with the signal voltage applied to give a vertical deflection. Used primarily to measure the time of occurrence (in radar, the range) of the signal, as well as its shape and strength. There are many modifications of this basic display type (see types K, L, M, N, O, R).



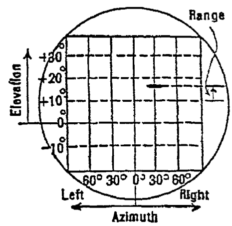
Type B

An intensity-modulated (radar) display in which the horizontal sweep is synchronized with the antenna direction, and the vertical sweep is a linear time base repeated with each transmitted pulse. The received signal is applied to the control grid or to the cathode of the CRT in such a polarity as to brighten the screen. The display has a large distortion from a true map when the range sweep starts with the transmitted pulse and a large azimuth sector is displayed. By delaying the range sweep, and by simultaneous proper control of the azimuth expansion, a good approximation to a true map may be obtained, over a limited region. Such variations are called "delayed B-scans" or "micro-B scans."



Type C

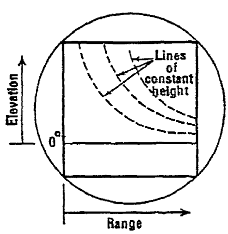
An intensity-modulated display, with azimuth as the horizontal coordinate and elevation as the vertical. This is the television type of scan. In radar, principally used in aircraft for location of other aircraft. Gives no indication of range. This type of indication is very bad from the signal-to-noise ratio standpoint, unless "range-gating" of the signal (from information obtained from a type B display, for example) is possible. Number of horizontal scanning lines used in radar is quite small, and signal appears on several scans because of wide angle of transmitted beam.



Type D

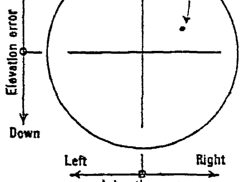
A modification of type C indication, for airborne radar, giving crude elevation information (widely separated elevation lines), but also yielding range information. As any line is scanned slowly from left to right, a short range sweep extending from one elevation line to the next is simultaneously applied. Thus, for the signal shown, the pertinent data are: azimuth, 30° to right; elevation, + 10°; range, 2 miles (assuming distance between elevation lines is equal to 3 miles).

Listed for record purposes only; never used in production.



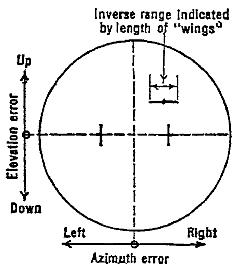
Type E

Similar to type B, but with range as the horizontal coordinate, and elevation angle as the vertical. Presents a distorted vertical cross-section, since lines of constant height are curved, as indicated. Chiefly used where elevation angle (and not height) is of importance.



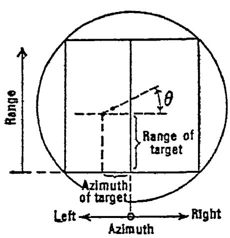
Type F

Sometimes called a "spot" error display. No sweep is employed, but the azimuth error of the signal is indicated by the horizontal coordinate and the elevation error as the vertical.



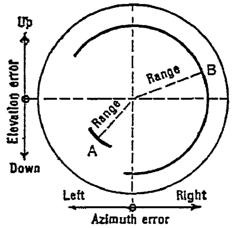
Type G

Sometimes called a "spot with wings" display. Similar to type F, with spot extended to a line whose length is inversely proportional to range to the target. Used mainly for gun-laying, the correct direction and range being indicated when the signal just fits between the two short vertical reference lines.



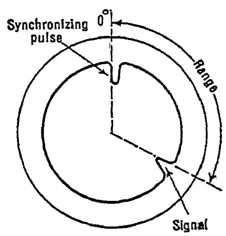
Type H

Known also as the "double dot" display, indicating range, bearing, and elevation. A modified type B display, in which alternate range sweeps are displaced slightly horizontally, thus producing two "dots" from a single echo. The position of the left dot is that which would be indicated by the type B display. The right dot is displaced vertically by an amount such that the angle θ is roughly proportional to the elevation of the target.



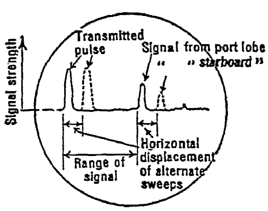
Type I

An intensity-modulated radar display, sometimes called a "broken circle" display, used with a conical scanning antenna. Each echo is represented by an arc of a circle whose radius represents range, and for which the shortness of the arc indicates the error in pointing at the target. Thus the arc A represents a target far to the left and below, while B is a target just a little above and to the right. A target dead ahead would be represented by a complete circle.



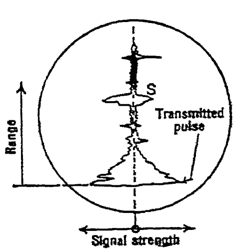
Type J

Essentially a type A display bent into a circle. Useful for accurate time or range measurement when a continuously running (crystal or other) oscillator can be used. The circular trace is obtained by applying two sinusoids from the same source to the two pairs of deflection plates, one sinusoid being 90° out of phase with the other. The synchronizing pulse which occurs at 0° also initiates the phenomenon to be observed, and the time to be measured, for example the range of a radar echo or the delay in a network is obtained from the angular position of the observed signal. The radial deflection is obtained by applying the signal voltage to a central electrode of the CRT (type 3DP1). The sweep may be delayed so that zero time occurs at 0° on some previous rotation of the sweep.



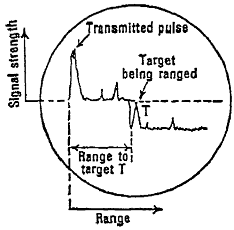
Type K

A radar display used with systems sending out two beams in slightly different directions. The type A sweeps from the two lobes are displaced slightly horizontally. By adjusting the angular position of the transmitting antennas until the signal height of the two corresponding displaced signals is the same, bearing or elevation information as well as range may be obtained.



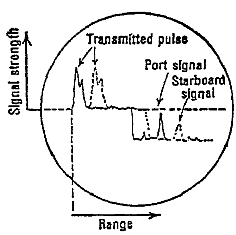
Type L

Sometimes called "back-to-back" A displays. Used with two-lobe radar systems, the signal *S* would indicate a target to the *left*. Changing the direction of the antennas until the opposite signals are of equal strength gives bearing or elevation of target as well as range.



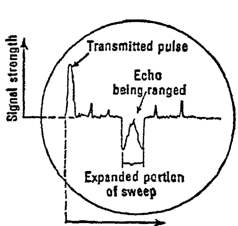
Type M

A modification of the basic type A display, in which a calibrated variable delayed "step" voltage is applied to one of the vertical deflection plates, in order to give more accurate range information. In practice, the step is shifted along the display until the signal to be ranged comes just at the edge of the step.



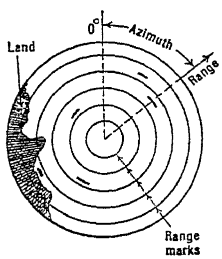
Type N

Combines features of types K and M in order to give more accurate range and bearing or elevation information.



Type O

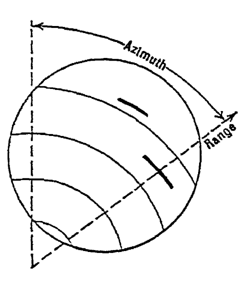
Sometimes designated as a "type A with a notch" display. Useful for accurate ranging on an echo. The horizontal sweep speed is increased for a short time interval, the beginning of the interval being adjustable and accurately calibrated. The signal under observation is placed in the center of this expanded portion of the sweep, whereupon the reading of the delay dial gives the range directly.



Type P or PPI

The most widely used intensity-modulated radar display, called plan position indication (PPI), in which the range of the echo is given by its distance from the center of the display, and its bearing by the angular position. The display may be either stabilized, so that north is at 0°, or unstabilized, in which case 0° represents the ship's or plane's heading. To produce this display, a radial sweep starting from the center at the instant of the transmitted pulse moves outward in the direction in which the antenna is pointing at the moment. The rotating sweep may be produced by a linear sawtooth current in a single-axis deflection coil mounted around the neck of a magnetic deflection CRT, the coil being rotated in synchronism with the antenna. Alternatively, a stationary two-axis coil may be used, with x- and y-axis sawtooth currents whose amplitudes are at any given time proportional to cos θ and to sin θ respectively, θ being the direction in which the antenna is pointing. Electrostatic deflection CRT may also be used, sawtooth voltages similarly modulated being applied to the two pairs of deflection plates.

Two modifications of the centered PPI are: (*a*) the "open-center PPI," in which increased accuracy in determining azimuth at close ranges can be obtained by starting the radial sweeps from a circle of arbitrary radius, and (*b*) the "delayed PPI," in which increased accuracy in determining the range of distant objects results from delaying the beginning of each sweep from the center by an arbitrary known time.

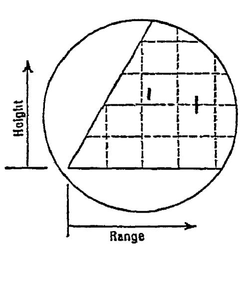


Off-center PPI or Sector Display

The off-center PPI is a variation of the centered PPI described above, in which the center of the pattern may be arbitrarily displaced in any direction, and the sweep expanded, so that any desired portion of the entire area within the range of the radar set can be made to fill the entire screen of the CRT. Such a display affords the possibility of greatly increased resolution while still preserving the undistorted map feature of the PPI. This is accomplished in the case of a rotating coil PPI by the superposition of a steady deflecting field upon the rotating sweep field. In the fixed x- and y-axis coil system, the same result may be attained by proper control of the delay and sweep speeds of the two components. The resulting display is often called a "sector display," although it may not be a complete sector if the origin of the pattern is not on the CRT screen.

Type R

The designation R (for range) applies to a type A display in which a fast linear sweep is employed, the start of the sweep being delayed by a precision calibrated delay circuit. Precision electronic time markers may also be used. The R display is generally used in conjunction with a standard type A display in which the time interval covered by the R sweep is indicated on the A sweep by intensifying the corresponding time interval. Both sweeps may be displayed on a single CRT at the same time by using electronic switching of the two sweeps and vertical displacement of the two traces.



Type RHI

This display is used to indicate the range and height of aircraft, hence the symbol RHI. Since the vertical distance (height) to be measured is in this case usually much less than the horizontal distance (range), the useful area of the CRT face can be greatly increased by expanding the vertical dimension. The display is usually obtained by applying simultaneous sawtooth currents to the x- and y-deflection coils of the CRT, such that $i_x = A \cos \theta \cdot t$ and $i_y = nA \sin \theta \cdot t$, where θ is the angle of elevation of the antenna, and n is the vertical expansion factor, usually between 5 and 10. In practice, the range sweep is usually simplified to $i_x = A \cdot t$, the resulting distortion being unimportant. Lines of constant height are straight and horizontal. The range sweep may be delayed for increased range resolution at large distances.

SECTION 16

SOUND-REPRODUCTION SYSTEMS

SOUND-REPRODUCTION SYSTEMS

Sound energy is transformed into some other form and retransformed back into sound for one of three general purposes: it may be desired to reproduce the sound instantly but at some other location, as in telephony or broadcasting; it may be desired to reproduce the sound at some subsequent time, as in phonographs and sound pictures; or it may be desired to reproduce the sound instantly and at the same location, but with increased energy content, as in public-address systems. The discussion herein is limited to reproduction by electric means, or at least partially by electric means, although other means are sometimes used (e.g., acoustic phonographs).

An electrical sound-reproduction system consists always of a microphone, or device to translate sound energy into electric energy; an amplifier and associated control equipment, to control the loudness of the sound; and a loudspeaker, or device to translate electric energy into sound energy. Similar considerations govern the choice and installation of these devices in all systems.

In addition, in phonographs and sound pictures there are devices to record the sound permanently and to effect the subsequent reproduction of the sound, as well as means to synchronize the recording and reproducing machines accurately. In telephone and radio broadcast systems a transmission link is inserted in the circuit; also in telephony a complicated switching system is included to permit any subscriber to talk privately to any other subscriber. (See Section 17).

AUDIO FACILITIES FOR SOUND SYSTEMS

By Howard A. Chinn

1. SOUND STUDIOS

Sound-reproduction systems in common use employ a single channel for transmission and for reproduction. Regardless of how many microphones are utilized for a given pick-up, their outputs are ultimately blended into one audio channel. This results in a single-channel system which imposes problems of a special nature upon the design of studios and sets used for sound broadcasting or for sound recording.

When a person listens to sound directly, the listening is done with both ears, or binaurally. *Binaural* hearing is indispensable for the localization of sound and for sound perspective. When listening to music in a concert hall, for example, one can largely ignore disturbing noises, provided they originate in a direction that is different from that of the desired sound and are not of unreasonable intensity. Furthermore, it is even possible to concentrate upon a particular section of the orchestra (for example, the strings in preference to the brass, or vice versa) and to ignore sounds that may be of no interest, or perhaps even unpleasant.

In listening to a single-channel system, on the other hand, nearly all sense of location of the sound is lost, as well as its extension in space. (A simple but effective demonstration of this is to block one ear, while conversing in a noisy room. The ability to continue to understand the speaker will be greatly impaired, if not entirely destroyed, because of the inability to exclude unwanted sound.) This same limitation exists when sound is transmitted over a single channel and reproduced by a loudspeaker system. Even though both ears are used for listening to the loudspeaker, the sound comes from a single source (multiple-unit loudspeakers, such as are sometimes required on wide-range, high-fidelity systems, do not change the effect) and the ability to discriminate against undesirable sounds is almost completely lacking. Hence, the sound that is heard from radio or television broadcasting, and in the cinema theater, is lacking in sound perspective. The design of studios for sound-reproduction purposes must be undertaken with these considerations in mind.

By combining acoustically correct architecture with suitable microphone pick-up and blending techniques, a single-channel sound system will produce pleasing results. However, the acoustical design problems are considerably more severe than those encountered in ordinary architectural work. Since the ability to localize sound is lacking, it is necessary to remove all sources of confusion or interference incident to monaural listening. This may be done by (a) proper control of reverberation, (b) the elimination of echoes, and (c) the elimination of disturbing noise. The first two items involve problems in acoustical treatment; the last one, problems of sound isolation—the two are not necessarily related.

OPTIMUM REVERBERATION TIME. Experience has shown that the most satisfactory reverberation time for broadcasting or sound-recording studios is less than that which is considered optimum for binaural listening. (See Section 12.) However, the reduction should be no more than is required to obtain the desired result. An excessive reduction in reverberation time will result in the elimination of the reverberant characteristics normally associated with the type of sound being reproduced. For example, the timbre or quality of a church organ being played in an acoustically dead room would sound as if it were out-of-doors. The reverberant quality associated with church music would be completely missing. The proper amount of reverberation must, therefore, be retained to create the desired psychological effect.

The use of studios that are acoustically too dead results in another undesirable effect. Musical organizations performing in an environment of this type are handicapped in several ways. The performer, discovering that his instrument does not produce the sound intensity to which he is accustomed, generally tries to produce the normal amount of sound, with the result that he concludes the studio is a "difficult" one in which to work. In addition, the performer may find that it requires special attention to hear the other members of his group in order to keep in time and in tune. Finally, the listener gains the impression that the orchestra is a much smaller aggregation than it actually is.

The type of the production dictates the optimum reverberation time for any given studio. In general, studios intended for speech or dramatic productions are less reverberant than those used for musical presentations. It is evident, therefore, that no one optimum will cover all situations. Furthermore, since maximum studio usage requires that a given unit be capable of accommodating all types of productions, some means of adjusting the acoustics is not only desirable but actually essential for full flexibility. In Fig. 1 there are shown recommended minimum and maximum reverberation times, at

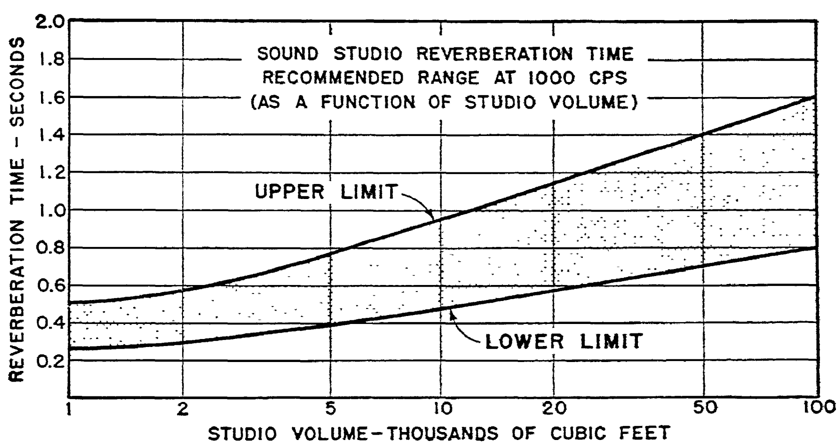


FIG. 1. Recommended Range of Sound Studio Reverberation Time (at 1000 Cycles)

1000 cycles, as a function of studio volume. The most reverberant condition shown is suitable for symphonic orchestras, church organ music, and other musical productions normally associated with large halls; the least reverberant condition is suitable for dramatic productions and musical programs where no "room tone" is desired.

It is desirable that studios be designed to provide a variation of reverberation time between the two limits shown in Fig. 1. This amount of flexibility makes it possible to adapt a studio for practically any kind of production. The advantages of being able to do this are obvious. The maximum load factor for every unit is assured without placing groups in studios having unsuitable acoustical characteristics. This feature alone will justify many times over the additional design problems that variable acoustics entails. Since a studio with variable acoustics can be used for all types of productions, fewer studio units are needed in any given group. The results are not only greater performer satisfaction but also a smaller initial investment and a smaller operating cost for the studio group as a whole.

VARIATION OF REVERBERATION TIME WITH FREQUENCY. In addition to the reverberation time at a specific mid-range frequency, the manner in which the reverberation time varies with frequency is important. Various investigators have discussed the shape of the reverberation characteristic from theoretical, subjective, experience, and environment viewpoints. Their findings are not conclusive, however. Furthermore, practically all were based upon binaural listening conditions.

Experience gained from broadcasting and recording studio applications seems to indicate that, except in the larger studios (i.e., those of about 50,000 cu ft or more) the reverberation time should be essentially independent of frequency (in practice, however, the humidity often limits the reverberation time at high frequencies). In the very large studios some increase in the reverberation time at the low frequencies is often beneficial for the type of productions for which such studios are used.

In a studio having variable acoustics, provisions may also be made to obtain variations in the shape of the reverberation-frequency characteristic. By properly locating various types of acoustical material on the absorbing side of adjustable panels (or on the wall back of movable panels), it is possible to change the shape of the characteristic, within limits. Variation, at the low frequencies, from a flat characteristic to a rising or falling one may often prove useful. A change of 20–30 per cent above a flat characteristic at 100 cycles to a like amount below should prove entirely adequate.

SOUND DECAY RATE. The manner in which sound decays in a studio is as important as the reverberation time, or even of greater importance. In some studios, sound does not decay logarithmically as geometrical acoustics assumes. Experience indicates, however, that a smooth logarithmic decay of sound results in the most acceptable type of studio. A decay curve of this nature insures the absence of discrete echoes and eliminates one of the sources of confusion to monaural listening already mentioned. For good acoustics it is generally considered desirable to have a constant decay rate for at least the first 30 or 40 db.

The desired type of sound decay can be obtained by creating a diffuse distribution of the sound. This can be done in a number of ways, such as employing a random distribution of the sound-absorbing materials, by the use of serrated walls and ceilings, or by substituting curvilinear surfaces for flat ones. It can be shown, theoretically, that curved panels, as contrasted to flat serrated panels, increase the area over which sound wave is dispersed. Likewise, it follows, that multicurved surfaces are an improvement over cylindrical ones. However, it is not yet generally realized that from a sound pickup viewpoint the diffusion of sound in an enclosure can be carried too far. Experience indicates that an entirely adequate amount of sound diffusion can be readily obtained with serrated flat surfaces. Consequently, there is little need to resort to the more complicated (from a construction viewpoint) curvilinear surfaces, except as may be deemed desirable from an esthetic point of view.

ACOUSTICAL NOISE LEVELS. The residual noise in studios, from all sources, should be as low as possible. However, there is a lower limit below which it is not justifiable to reduce the noise from either a theoretical or economical viewpoint. The maximum sound intensities normally encountered in studios at the normal microphone pick-up position (relatively close to the speaker for speech but relatively distant for music) range from about 75 db (above the acoustical reference level of 10^{-16} watt per cm^2) for speech to 95 db for music. The listener, on the other hand, indicates that noises that are 50 to 60 db below the signal are either unobjectionable or not detectable. Taking these considerations into account, together with the economic ones, it is evident that a residual-noise level 25 db above the reference level is quite satisfactory. Sound isolation methods are discussed in detail in Section 12.

BIBLIOGRAPHY

Morse, P. M., and R. H. Bolt, Sound Waves in Rooms, *Rev. Modern Physics*, April 1944, p. 69. (Includes an excellent bibliography.)

Articles in the *Journal of the Acoustical Society of America*, *Journal of the Society of Motion Picture Engineers*.

2. MICROPHONES

The types of microphones most commonly employed for program pick-up purposes are: condenser, moving-coil (or dynamic), velocity-actuated ribbon, crystal (piezoelectric), and combination units. These last microphones usually combine a velocity- and a pressure-actuated ribbon or velocity-actuated ribbon and a moving coil into a single unit. The phasing between the two units of the assembly is arranged so as to obtain a directional effect, such as a unidirectional or a cardioid pattern. The characteristics of various microphones are discussed in detail in Section 13.

The output level of wide-range, high-quality microphones is extremely low—so low, in fact, that the signal-to-thermal-noise ratio existing at the output terminals of the microphone often determines the overall performance of the system. The sound intensities existing at the microphone position for the usual type of program productions is such that a net insertion gain of 50 to 60 db is usually required to raise the output level to 0 vu, the standard reference level (see below). Still greater gain is required, of course, to raise the level sufficiently for transmission over program circuits, for operating loudspeakers and recorders, or for modulating a transmitter.

The output impedance of condenser and crystal microphones is very high, and, as a rule, these units are operated directly into the grid of a tube. On the other hand, the output impedance of the moving-coil, ribbon, and combination microphones has ranged from a few ohms to several thousand ohms, depending upon the type. Such a variation greatly restricted the universal use of various kinds of microphones with different equipment. The situation has now been recognized, and microphones intended for broadcasting purposes have been standardized at 150 ohms. In most cases, a transformer must be supplied as an integral part of the microphone in order to provide the standard output impedance.

MICROPHONE PLACEMENT. In determining the proper placement of a microphone for a given type of program it is necessary to take into consideration the directional properties of the particular microphone being used as regards both its amplitude and its frequency response characteristics. Each type of microphone has a directional characteristic peculiar to that type, and both the horizontal and the vertical plane directional pattern of the microphone must be considered in connection with its application and placement. Some microphones are unidirectional, others bidirectional, and still others nondirectional in the horizontal plane. These properties are extremely important and useful for discriminating against undesired sources of sound and for obtaining a desired relation between sounds from different sources.

Figure 2 illustrates the placement of two bidirectional microphones which, for the purposes of illustration, are assumed to have a figure 8 directional pattern in the horizontal plane. The performance being picked up is assumed to consist of a chorus, a soloist, and an accompanying orchestra. Microphone A is used for the chorus and soloist pick-up, the artists being grouped on both sides of the instrument. The orientation of this microphone is such that the null point or direction of minimum pick-up is towards the orchestra. Microphone A, therefore, is actuated primarily by the chorus and soloist and picks up very little of the orchestra music. Microphone B, on the other hand, is so located as to derive its major source of energy from the orchestra, and its null point is towards the singers. The electrical outputs of these microphones, one of which consists primarily of the voices and the other of the musical accompaniment, are then properly combined or "mixed" so as to obtain the desired balance. The technician responsible for this operation has control over the relative magnitudes of the singers' voices and the music. He can adjust the relation between the two and also the overall amplitude to obtain any desired result. Still other microphones may be used for picking up sound effects, announcements, audience or crowd noise, etc. By the proper placement each microphone can more or less be confined to the pick-up of its assigned source of sound.

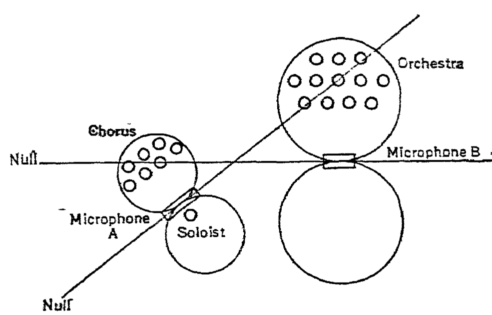


FIG. 2. An Illustration of the Placement of Bidirectional Microphones to Achieve Control of Sound Pickup

Microphones having relatively sharp directional characteristics are sometimes used for outdoor or long-distance sound pick-ups. These devices are particularly good for confining the source of pick-up of sounds at a large outdoor gathering to a particularly interesting part of the crowd such as a cheering section or a band of musicians. In television pick-ups they provide a means for keeping the microphone out of the camera angle.

The use of more than one microphone for the pick-up of a given performer or group of performers working as a unit (orchestra, chorus, etc.) is generally to be avoided. Serious frequency and delay distortion is likely to result if more than one instrument is employed for the pick-up because under these circumstances each microphone will be a different distance away from a given source of sound. As a result the sound waves do not reach each microphone at the same instant, and the combined outputs of the microphones will result in complete or partial reinforcement or cancelation, depending upon the resultant phase relationships. Furthermore, the phase relation is dependent upon the frequency of the sound source and does not, therefore, remain a constant quantity for all sounds.

In practice it is sometimes advisable, however, to countenance these potential sources of distortion and use more than one microphone for the pick-up of a given group of performers. For example, an orchestra may have a string choir too small for good balance. Under such circumstances, supplementing the main microphone by a strategically placed and judiciously used secondary microphone may enhance the overall balance. It is important to note, in an operation of this type, that the contribution of the additional microphones must always be maintained at a low level.

The frequency-response characteristic of some microphones varies with the angle of incidence of the sound wave upon the instrument. Microphones that exhibit this characteristic usually have a decreasing response to the higher frequencies as the source of sound moves around from the front to the side of the instrument. In placing such microphones this characteristic must be taken into consideration. The unidirectional dynamic and the condenser microphones are of this type. The velocity, non-directional dynamic, and crystal microphones, on the other hand, maintain the same relation between the high and low frequencies at all angles of incidence.

BIBLIOGRAPHY

Hopper, F. L., Characteristics of Modern Microphones for Sound Recording, *J. Soc. Motion Picture Eng.*, September 1939, p. 278.

Marshall, R. N., and W. R. Harry, A Cardioid Directional Microphone, *J. Soc. Motion Picture Eng.*, September 1939, p. 254.

Marshall, R. N., and F. F. Romanow, A Non-directional Microphone, *Bell Sys. Tech. J.*, July 1936, p. 405.

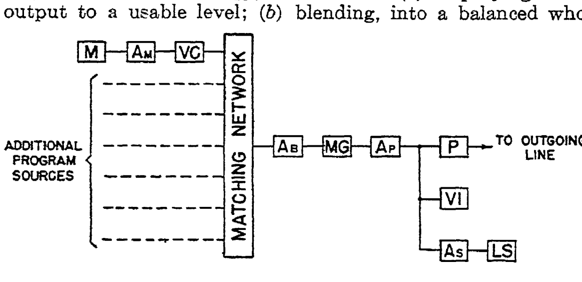
Olson, H. F., *Elements of Acoustical Engineering.* Van Nostrand (1940).

Olson, H. F., The Ribbon Microphone, *Proc. I.R.E.*, May 1933, p. 655.

3. AUDIO AMPLIFIERS AND CONTROL EQUIPMENT

A complete sound system entails, in addition to microphones for converting sound energy into electrical energy, facilities for: (*a*) amplifying the exceedingly low microphone output to a usable level; (*b*) blending, into a balanced whole, program elements from several channels (known as "mixing"); (*c*) adjusting the balanced program to the desired transmission level *without* altering the balance; (*d*) aurally and visually monitoring the transmission.

These elements are found in audio systems used for broadcasting, sound recording, and public-address applications. Figure 3 illustrates, in block diagram form, a typical audio system incorporating the elements listed above. The components of the complete system are described below in the order in which they appear in the block diagram.



LEGEND

AB	BOOSTER AMPLIFIER	M	MICROPHONE
AM	MICROPHONE PRE-AMPLIFIER	MG	MASTER GAIN CONTROL
AP	PROGRAM AMPLIFIER	P	ATTENUATION PAD
As	MONITOR AMPLIFIER	VC	MIXER GAIN CONTROL
LS	LOUDSPEAKER	VI	VOLUME INDICATOR

FIG. 3. Simplified Block Diagram of Typical Sound System Studio Audio Facilities

PRELIMINARY AMPLIFIERS. It is common practice in reproduction systems to have a mixer volume control associated with each microphone or other source of pick-up. As already noted, however, the output level of microphones is so low that any attenuation, prior to amplification of the signal, would degrade the signal-to-noise ratio of the system (the cause of the noise at this point being largely thermal). Therefore, before any volume controlling can be done, it is necessary to raise the level of the signal well above the thermal-noise level. In fact, one of the cardinal principles of good audio system design is never to permit the signal level to fall below the value existing at the output terminals of the microphone. It is therefore common practice to insert a preliminary amplifier between the microphone and the mixer volume control (or any other circuit element such as a sound-effects filter, dialogue equalizer, etc.). By using this arrangement the mixer control is introduced into a circuit at a point where the level of the microphone output has been raised considerably above the thermal-noise level of the circuit. As a result, when attenuation is introduced for mixing purposes, the signal-to-noise ratio is not degraded.

Preliminary amplifiers usually consist of one or two stages of amplification having an overall gain in the neighborhood of 30 to 40 db. The amplifier must be carefully designed because any noise originating in it will be amplified by the following amplifier stages. It is therefore imperative that amplifier noises such as microphonics and hum background be practically non-existent in the preliminary amplifier. The input transformer of the preamplifier, if one is used, is a particularly susceptible point for the pick-up of stray, unwanted, interfering fields. This unit must be very carefully shielded both electrostatically and electromagnetically.

Low-impedance microphones, such as the dynamic and velocity types, are connected to the input of the first amplifier tube by means of a suitable transformer. However, microphones of this type are basically voltage-generating devices; consequently their output impedance (150 ohms in broadcasting practice) is *not* matched to the input impedance of the preliminary amplifier. Rather, the preliminary amplifier is designed to have a high input impedance, thereby realizing as much of the open-circuit voltage of the microphone as possible. The input transformer is normally designed with as high a step-up ratio as commensurable with the required response-frequency characteristic. The leads connecting the microphone to the preamplifier input transformer may be several hundred feet long without seriously impairing the performance of the device.

The condenser and crystal types of microphones are connected directly to the input of the amplifier tube through a suitable network of resistors and condensers. In this case it is desirable that the leads connecting the pick-up device to the amplifier tube be of very low electrostatic capacitance. (Connecting cable capacitance attenuates the microphone's output voltage but does *not* impair its response as a function of frequency.) It is common practice to make these leads very short by building the preamplifier into the mounting which houses the microphone head. In the floor-stand type of mounting this amplifier is sometimes built into the base of the stand.

In the design of preliminary amplifiers care must be exercised that the output stage has adequate power-handling capacity. Since no gain control is ever inserted between the microphone and the preliminary amplifier, the input to the amplifier will vary over wide limits. The output level of the usual microphone under the average conditions met in practice has already been noted. However, if a microphone is placed close to a musical instrument that is being played very loudly (or near another source of loud sound) its output level may be as much as 20 db higher than that "normally" obtained. The output stage of preliminary amplifiers must be capable of handling without overloading the output level resulting from this input level. The output of the preamplifier is usually matched to the mixer circuit impedance by means of a suitable output transformer. For broadcasting applications the standard output impedances are 600 and 150 ohms.

MIXER VOLUME CONTROLS. A mixer circuit is an arrangement of volume controls for combining into one program channel, in any desired proportions, program elements of several channels. The multiple microphone, transition, and fading effects, which contribute so much to program continuity, are obtained with mixer controls. For instance, separate microphones may be used for a soloist and for the accompanying orchestra (Fig. 2) and the outputs combined to form a balanced whole. Since the gain of each microphone channel may be regulated independently of the others, a method of controlling the balance between various pick-ups is provided which does not impair the quality of the individual sources of sound.

A mixer is also used to "fade down" a musical transmission so that announcements or talks may be superimposed. All these effects contribute a degree of smoothness to a program which would otherwise be impossible. Because of their use these controls are sometimes known as "faders."

The mixer control is, in effect, a variable-resistance attenuation pad, usually of a T structure, having a constant or nearly constant input and output impedance. The device is capable of providing attenuation over a range from 0 to about 120 db. The attenuation generally varies uniformly with the angle of rotation of the control knob throughout the range from 0 to approximately 50 db. From this point to maximum attenuation the increase is very rapid, occurring in perhaps one-tenth of the full arc of rotation. For broadcasting applications the standard mixer control impedances are 600 or 150 ohms.

MIXER MATCHING NETWORK. A mixer matching network combines the output of a number of mixer controls into a single channel while maintaining correct impedance relations. The matching network provides the proper load impedance for the individual mixer controls and also the desired source impedance for the following circuit element.

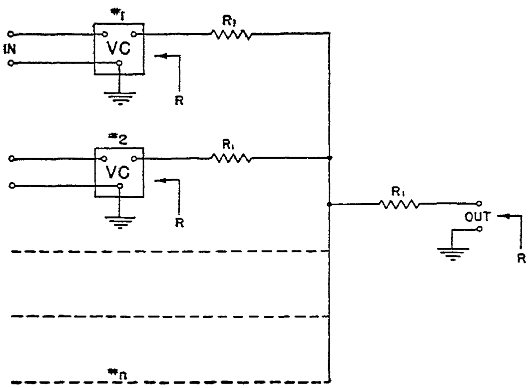


FIG. 4*a*. Differential Matching Network Having Like Input and Output Impedance

One of the simplest forms of matching networks is shown in Fig. 4*a*. In the general case of n mixer positions, the value of the building-out resistors, R_1, is

$$R_1 = \frac{R(n-1)}{n+1}$$

Where the input and output impedance of the network are assumed to be alike and equal to R.

The loss between the output of any given mixer control and the matching network load is

$$\text{db loss} = 20 \log_{10} n$$

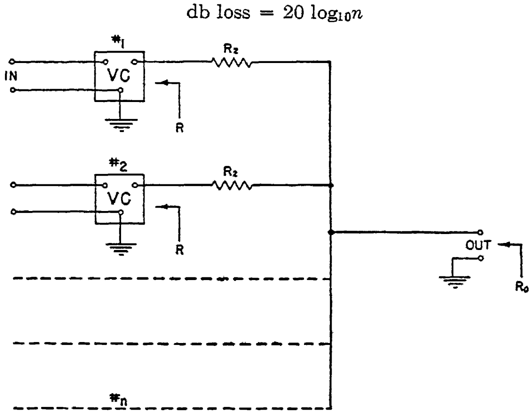


FIG. 4*b*. Minimum-loss Differential Matching Network. Input and output impedances are *not* alike.

If the requirements of like input and output impedance are waived, a lower loss network can be used, Fig. 4*b*. Here, the building resistor value R_2 is

$$R_2 = \frac{R(n-1)}{n}$$

The output impedance R_0 is

$$R_0 = \frac{R(2n-1)}{n^2}$$

The loss of the network is

$$\text{db loss} = 10 \log_{10} (2n-1)$$

In this case the output impedance can be restored to the same value as the input impedance (or to any other value) by means of a matching transformer.

BOOSTER AMPLIFIER. The amount of attenuation introduced in an audio system by the mixer controls and the associated matching network is often so great that amplification must be supplied before further volume controlling (see Master Volume Control, next paragraph) can be effected. The amplifier used for this purpose is termed a booster amplifier. It usually employs one or two stages of amplification, has a gain of 30 to 40 db, and is designed to operate from a source impedance and into a load impedance of finite value (e.g., 600 and 150 ohms in broadcast service).

Whether a booster amplifier is required between the mixer circuit and the master gain control is determined by the signal levels existing for normal input levels and normal settings of the controls. Its use is indicated if, by its omission, the signal would fall below the level existing at the output terminals of the microphone.

MASTER VOLUME CONTROL. The master volume control or "gain" control is "master" over the output of the individual mixer controls. After these mixer controls are adjusted to obtain the proper relation between the various parts of the performance the resultant overall volume of the program material may be adjusted to the desired level by means of the master gain control without affecting the balance that exists. This device also permits the properly balanced performance to be faded in or out while the same relation is maintained between the individual parts of the program. The master gain control is similar in construction and operation to the mixer volume controls.

PROGRAM AMPLIFIER. The purpose of the program amplifier is to bring the level of the studio output up to the point necessary to permit its being fed directly into the audio amplifier stage of a transmitter, into the program line connecting the studio with a transmitter, into a loudspeaker amplifier or a recording system. At a network key station, the output of the program amplifier is fed into a bus for distribution to the proper network or networks.

This amplifier follows the preamplifiers, mixers, and master gain control to amplify the output of the microphones or other program source further. Although one preamplifier is needed for each microphone in use, only one program amplifier is necessary for a given studio channel. The amplification obtainable in these units is generally in the vicinity of 50–60 db. The amplifiers are usually capable of an output level of approximately 250 milliwatts without serious overloading. The input impedance of this amplifier is matched to the output impedance of the preceding master gain control by means of a suitable transformer. The output impedance is similarly matched to the load which follows this unit.

BIBLIOGRAPHY

Chinn, H. A., Broadcast Studio Audio Frequency System Design, *Proc. I.R.E.*, February 1939, p. 83.
Chinn, H. A., CBS Studio Control-Console and Control-room Design, *Proc. I.R.E.*, May 1946, p. 287.
Monroe, R. B., and Palmquist, C. A., Modern Design Features of CBS Studio Audio Facilities, *Proc. I.R.E.*, June 1948, p. 778.

4. MONITORING FACILITIES

AURAL MONITORING FACILITIES. A monitoring system consisting of a suitable amplifier and loudspeaker is a part of every complete sound system. These facilities provide a means for those responsible for the production of hearing the program exactly as it is being sent to the transmitter, the network, or the recorder. It is essential that the fidelity of the aural monitoring equipment be in keeping with the remainder of the broadcasting or recording system.

An output of 10 to 25 watts and an amplification of approximately 50 db are generally obtainable from the monitoring amplifier. The input impedance of the monitoring amplifier is usually very high in order that it may be bridged across the program circuit at a convenient point without affecting the level or impedance balance of the circuit to any appreciable extent. The output circuit is provided with a suitable transformer to effect an impedance match with the particular type of loudspeaker or speakers being employed.

Broadcasting and sound-recording control rooms often afford a limited amount of space for the installation of a monitoring loudspeaker, placing several special requirements upon the design of the unit. For instance, the directional properties of the loudspeaker must be such as to provide uniform coverage for all occupants of the control room who are concerned with the production in hand. This requirement usually entails some special arrangement to insure uniform distribution of the high frequencies. Furthermore, since the loudspeaker may be relatively close to the listener, multichannel loudspeakers (if used) must be especially arranged so that the separate sources of sound cannot be distinguished as such.

VOLUME INDICATOR FOR VISUAL MONITORING. A volume indicator is used in order to provide a precise, visual means of determining the volume level of the program material being transmitted by an audio system. The standard volume indicator consists of a copper oxide rectifier and a d-c indicating instrument. The characteristics of the rectifier, together with the dynamic, electrical, and other performance characteristics of the indicating instrument, are all carefully standardized. When calibrated and used in the prescribed way, the standard volume indicator gives an accurate indication of volume level. This is expressed as so many "vu" above (or below) reference volume—the number of vu being numerically equal to the number of decibels that the volume level is above (or below) reference volume.

REFERENCE VOLUME—VU. Reference volume is that level of program which causes the standard volume indicator, when calibrated and used in the prescribed way, to read 0 vu. By definition, the reading of the standard volume indicator shall be 0 vu when it is connected to a 600-ohm resistance in which there is flowing 1 milliwatt of sine-wave power or n vu when the calibrating power is n decibels above one milliwatt.

Reference volume, as applied to program material, should not be confused with the single-frequency power used to calibrate the volume indicator. Speech or program waves that result in a volume reading of 0 vu have instantaneous peaks of power many times 1 milliwatt and an average power which is a fraction of 1 milliwatt. In other words, reference volume is not 1 milliwatt, except in the special case of sine-wave measurements.

DBM VS. VU. For steady-state measurements a reading in "vu" denotes a specific single-frequency audio power; for dynamic program indications vu denotes only a "volume" level. This dual meaning of vu is avoided by use of the term "dbm" for all steady-state measurements. Using this terminology, a reading expressed in dbm indicates the *power level* of a steady *single-frequency* signal where the number of dbm is equal to the number of decibels above (or below) a reference power of 1 milliwatt. On the other hand, a reading in vu denotes a volume-level indication of a program signal. A vu reading can be made only on a standard volume indicator (since the dynamic characteristics are involved) whereas sine-wave power level measured with the standard volume indicator or with any other suitable a-c instrument can be expressed in dbm.

The practice of expressing measurement-signal levels in dbm, and of limiting vu to expression of dynamic volume levels, has certain advantages. Thus, dbm is a unit of finite audio power, whereas vu may be considered only a unit of volume level and, as mentioned above, has no connotation of finite power level.

However, it is necessary to establish a relation between the vu level to be used for program peaking and the dbm level to be used for performance measurements. It has been found that on typical program peaks reaching a given crest amplitude the standard volume indicator reaches an indication 8 to 14 db below that reached on a steady tone of the same crest amplitude. To take into account this 8- to 14-db difference in response, it is standard practice to require that performance requirements be met at a single-frequency test-tone level 10 db higher than the normal program level. This will reasonably insure that system performance is within standards under most operating conditions.

TALK-BACK EQUIPMENT. A "talk-back" system is generally employed in order to provide a means for communication between the control room and the studio during the course of rehearsals. This equipment consists of a microphone in the control room, a suitable amplifier, and a loudspeaker in the studio. By means of this equipment the technician or the production man in the control room may talk to and direct the performers in the studio during rehearsals.

The talk-back equipment is interlocked with the regular studio equipment so that the studio microphones and the control-room monitor speaker are turned off whenever the talk-back equipment is energized for use. This prevents the generation of an acoustic feedback because of coupling between the input and the output circuits of the amplifier systems. The regular preamplifiers, studio channel amplifier, and monitor amplifier associated with the studio are sometimes employed, by suitable switching means, for the talk-back service. Inasmuch as both services do not function simultaneously, this arrangement results in an economical use of the equipment already available. The switching complications, however, often do not justify this practice. Under these circumstances separate or partly separate talk-back equipment is used.

BIBLIOGRAPHY

Chinn, H. A., D. K. Gannett, R. M. Morris, The New Standard Volume Indicator and Reference Level, *Proc. I.R.E.*, Jan. 1, 1940; also *Bell Sys. Tech. J.*, Jan. 1, 1940.
Chinn, H. A., DBM vs. VU, *Audio Engineering*, March 1948, p. 28.

ELECTROACOUSTIC EQUIPMENT

By R. J. Kowalski

5. AUDITORIUM ACOUSTICS

When an auditorium is designed for use with a sound-reproducing system one of the most important design considerations is the acoustic characteristics. (See also Section 12). The sound emanating from the stage loudspeakers reaches the listener only after being influenced by the acoustic conditions of the auditorium. In view of the technical perfection of modern sound-reproducing equipment, it is frequently these acoustic conditions that determine whether the sound reaches the ear of the listener with all its original naturalness and realism or whether it is distorted, unnatural, and wholly unsatisfactory to the listener.

The most common acoustic defects encountered in auditoriums are reverberation, echo, resonance, poor distribution (loud and dead spots), and noise.

Optimum reverberation times for auditoriums of various sizes are shown in Fig. 1. It will be noted that as the volume increases the allowable reverberation time also increases.

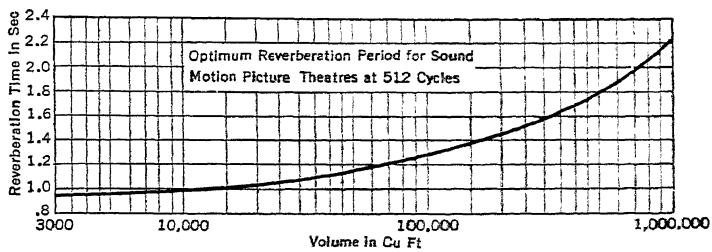


FIG. 1. Optimum Reverberation Times for Auditoriums

It will also be noted that the optimum reverberation time recommended is from 10 to 20 per cent below the usual values for an auditorium that uses live talent without a sound-reinforcing system. This is because there is an abundance of sound energy and hence reverberation is not necessary to augment loudness. In fact some recent thinking favors overtreating an auditorium with sound-absorptive material to eliminate most acoustic defects and then artificially introducing the proper amount of reverberation in the sound-reinforcing system by means of reverberation or echo chambers. These can be designed to give any desired amount of reverberation and can be quickly and cheaply changed to meet different conditions.

It will be noted in Fig. 2 that at the lower frequencies longer reverberation periods can be tolerated while at the higher frequencies the allowable period is somewhat reduced.

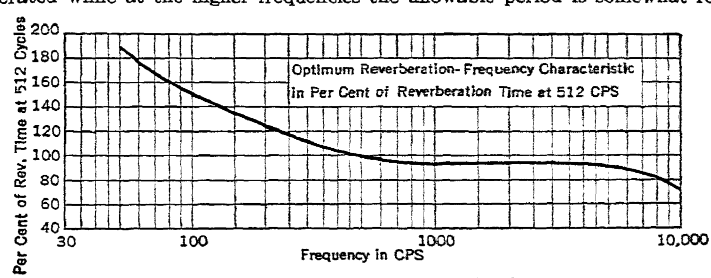


FIG. 2. Optimum Reverberation Time with Frequency

By referring to the table of absorption coefficients of building materials in Section 12 it will be noted that most materials have a different coefficient for different frequencies. Because of this, it is possible by proper selection of materials to arrive at a design which will give optimum reverberation time throughout the frequency range. Owing to the extension of the low-frequency and the high-frequency ranges of modern reproducing equipment great care should be observed in the choice of sound-absorbing materials and treatment.

Echo consists of a delayed repetition, sometimes several rapid repetitions, of the original sound. It is most often encountered in large auditoriums, particularly those with curved ceilings and walls and other surfaces sufficiently remote from the source of original sound to cause a definite time interval between the arrival of the original and reflected sounds at the listener's position. A multiple or flutter echo (several distinct repetitions) is often caused by parallel walls with smooth hard surfaces. With the extension of the high-frequency range of modern equipment the problem of echo and sound concentration was somewhat intensified because the high frequencies, or high-pitched notes, on account of their short wavelengths, are more easily reflected by small, smooth surfaces.

Echoes have much the same effect as reverberation in that they tend to blur speech and music. Echoes are eliminated by first localizing them and then applying light drapes or other sound-absorbing materials, or by breaking up the regularity of the offending surfaces by stepping, angling, etc., thereby dispersing or scattering the sound striking them. When it is necessary to add sound-absorbing material to correct for reverberation time, it is best to apply it first to the rear wall of the auditorium. Since the speakers are directed toward this wall it is usually the worst offender for echoes. If the side walls are parallel, the next best place to apply treatment material is on alternate panels of the side walls with the treated panels staggered so that no two untreated surfaces are opposite each other. This helps to eliminate flutter echo between walls. Flutter echo between ceiling and floor can best be avoided by specifying heavy carpeting with padding in the aisles and heavily upholstered seats.

The phenomenon of resonance, or the ability to vibrate best at certain frequencies, may occur either in structures or in the air in rooms. The effect is a build-up or overemphasis of certain frequencies. Structural resonance usually is not harmful unless the resonant body is mechanically coupled close to the source of sound. An offending object can usually be located quickly by connecting an audio-frequency oscillator to the input of the sound-reproducing system and varying the frequency until resonance is reached. By surveying the room while this frequency is maintained, the vibrating object can usually be located. The vibration can then be eliminated by changing the mounting or adding damping material. Resonance in air chambers is usually encountered in small, highly reverberant rooms; it rarely is a problem in the main body of a large auditorium, though it may give trouble in smaller sections such as the stage, which usually has hard, bare, parallel walls, or under a balcony, in alcoves, or in foyers. To eliminate such resonance conditions on the stage, as well as to help reduce reverberation on the stage, absorbing material should be draped in the region around the loudspeakers.

Poor distribution of sound (i.e., loud and dead spots due to the shape of the auditorium) can usually be overcome by the proper type and proper orientation of the loudspeakers. This problem is discussed further in article 6.

Noise may be defined as any unwanted sound. Noise is undesirable particularly because it has a masking and a frequency discriminating effect on the desirable sounds which, therefore, require added loudness or power to override the noise An auditorium in a noisy city location should have its outside walls insulated against the transmission of outside noises into the auditorium. See Section 12 for a full discussion of noise reduction.

The following recommendations will serve as a guide in designing an auditorium to get the best results from a modern sound-reproducing system:

1. All seats should be of the heavily upholstered type.
2. Heavily padded carpeting should be used in all aisles and corridors.
3. The rear wall, being potentially the greatest source of echo, should be lined with an efficient type of sound-absorbing material and/or sloped or otherwise shaped to direct the reflected sound to nearby audience or treated areas to prevent echo.
4. Surfaces with concave curvature should be avoided as much as possible. If such surfaces are necessary they should be broken up with smaller convex flutes to disperse the sound or heavily treated to absorb most of the incident sound.
5. Large unbroken surface areas, except when used for beneficial reflections, such as reflection from the splayed ceiling and walls at the proscenium, should be avoided.
6. Long narrow auditoriums, high ceilings, and excessively long and low balcony overhangs should be avoided.
7. The cubical content of the auditorium should be made as small as possible, compatible with the seating capacity and architectural design.
8. A rising slope in the orchestra floor should be used to give unobstructed "sound" lines as well as "sight" lines.
9. All auditorium walls should provide sufficient sound insulation to prevent extraneous noises from entering.
10. All machinery and ventilating noises should be isolated from the auditorium.

POWER REQUIREMENTS. In calculating the power requirements for a sound-reproducing system there are several factors to be taken into consideration. In an outdoor installation the listener gets only the direct sound from the speakers, while in a small room or auditorium the total energy reaching the listener is the sum of the direct energy and the reflected reverberant energy. Hence the total absorption of the auditorium must be known to determine the acoustic power required. It is also necessary to know what sound pressures it is desired to attain before calculating acoustic power. For high-fidelity reproduction of both sound and music it is generally assumed that sound pressures on peaks and loud passages will reach 20 dynes per sq cm, or 100 db above the threshold of audibility. Finally, since we are chiefly interested in the electrical power needed in the reproducing amplifiers, it is necessary to consider the conversion efficiency of the loudspeakers and the coverage efficiency, that is, the percentage of the total radiated sound energy actually distributed to the required area.

In making calculations for any given installation, the following formula is recommended:

$$P = 0.0009 \frac{S_F}{E_L} \left[\frac{A_0 S_T}{A_0 S_T + 4R_F S_F} \right] \text{ watts}$$

where S_F = floor or seating area in square feet.
E_L = loudspeaker efficiency expressed in fractions thus, 50 per cent = 0.50.
A_0 = average absorption coefficient of room.
S_T = total area of room surfaces in square feet.
R_F = reflection coefficient of seating area. This equals 1 minus the absorption coefficient $R_F = (1 - A_F)$.

The constant in this formula is calculated to give a sound pressure of 100 db. If less acoustic power is required, as in small rooms, the power as calculated by the formula may be decreased an equivalent number of decibels. For outdoor reproduction, the absorption is considered infinite, hence the factor in brackets goes to unity, simplifying the equation to $P = 0.0009 S_F/E_L$. Assuming an efficiency of 45 per cent, which is average for a modern directional loudspeaker of the type used in outdoor installations, the formula becomes $P = 0.002 S_F$. This indicates that, in order to develop sound pressures of 100 db without the benefit of reverberation, using a loudspeaker with an overall efficiency of 45 per cent, it is necessary to supply 2 watts of electrical energy for each 1000 sq ft of audience area.

6. LOUDSPEAKERS

Through the years, a good many different types of loudspeakers have been developed to convert electrical energy into acoustic energy (see Section 13). Among them were the head-phone-type magnetic-diaphragm units, the condenser speakers, the magnetic-armature speakers, and the moving-coil-type speakers. In recent years the moving coil or so-called dynamic speakers have proved the most efficient, and hence this is the only type now in common use. The magnetic field that the voice coil moves in is supplied by a field coil energized with direct current or by a permanent magnet. With the new magnetic alloys like Alnico it is possible to get flux densities in the permanent-magnet type that are as great as those obtained when using an electrically energized field. Hence the efficiencies of the two types are comparable. The permanent-magnet type is rapidly becoming the most popular, because, though it is a little more expensive to build, it saves the cost of an electrical field supply as well as considerable additional wiring, which is quite a factor when a good many speaker units are located at remote points. Aside from the type of field, dynamic speakers are classified in two general types: (1) the direct radiator type and (2) the horn type. The direct radiator type uses a large cone with a relatively small-diameter voice coil for a radiator. This unit is generally mounted on a flat baffle or in a cabinet. The horn-type unit has a relatively small diaphragm driven by a large-diameter coil. This unit is specifically designed to be used with a directional type of horn.

The direct radiator or cone type of speaker on a flat baffle is generally used in small spaces such as hotel guest rooms, school classrooms, or hospital ward rooms. Although this type of loudspeaker is frequently considered non-directional, it has a definite distribution angle, particularly at the higher audio frequencies. For a rough approximation at the most important frequency range, this angle may be considered to be 90°, that is, 45° off from either side of the central axis. Since the shape of the cone is symmetrical, this distribution angle is the same for both the horizontal and vertical planes. In determining the number of speaker units to place in a given room, not only the power-handling capacity but also the coverage must be considered. A floor plan of the room may be sketched to scale, and speakers may be located, and the coverage of each plotted.

If the distribution angle of one speaker does not cover at least 75 per cent of the total area, two or more speakers should be placed along the wall until this much of the area is covered. All speakers must be connected to operate in phase. When speakers of this type are used for incidental music such as in a hospital or restaurant it is preferable to increase the number of speakers and hence cut down the power per speaker to get the most uniform sound intensity throughout the room.

For larger installations in auditoriums or in unconventionally shaped rooms much more efficient distribution of sound energy can be obtained through the use of horn-type loudspeakers with directional horns. The more concentrated beam distributes the sound in the desired area and at the same time keeps it off walls and ceilings, hence helping to avoid serious echo effects. In de luxe installations where high-fidelity reproduction of music is desired best performance can be obtained by using a two-way speaker system with multicellular exponential high-frequency horns and a folded baffle for the low-frequency speakers.

In auditoriums or theaters, the location of the speakers for best illusion is usually above and to the front of the proscenium arch. If this location will not allow for the projection of sound into the rear of the orchestra floor under the balcony, it will be necessary to place additional speakers to the sides of the proscenium directed to cover this area. These speakers should be kept as high as possible and should be angled to get as little sound as possible into the front rows of seats. When locating speakers, caution should be used to see that the projected sound beam does not pass through the microphone pick-up area where it would cause acoustic feedback and hence limit the gain of the reinforcing system.

When more than a single speaker unit is used, it is essential that all units operate in phase. On better speakers, the phasing is carefully controlled in production and the terminals of the speaker are marked so it is only necessary to connect all similar terminals together to get proper phasing. If the speakers are not marked it is possible to energize the speaker fields and then apply a small d-c voltage to each voice coil. By reversing the polarity of the battery it is possible to determine to which terminal of the speaker the positive terminal of the battery should be connected to give a forward deflection of the speaker diaphragm. If these positive terminals are connected all the speakers will be properly phased.

7. AMPLIFIERS AND CONTROL EQUIPMENT

Before selecting the proper amplifier for a given installation it is necessary to determine the power requirements as described in article 5. Strict adherence to this formula in small auditoriums might indicate relatively low power requirements; however, the Research Council of the Academy of Motion Picture Arts and Sciences recommends a minimum of 10 watts of audio power for the smallest of theaters. Though the average power used in normal reproduction will be considerably below this figure, the reserve power will make it possible to reproduce peaks and loud passages without distortion. This minimum of 10 watts is recommended for auditoriums up to 500 seats. The recommended power rises approximately 10 watts for each additional 500 seats. In multiroomed installations like hotels, the total power requirements are found by adding up the power requirements of the individual rooms. Systems of this type frequently have a varying use factor for each room so that all rooms might not have to be supplied simultaneously. Although this use factor varies widely with different types of installations, the suggested values for hotels are: 100 per cent for a single channel installation, 90 per cent for two channels, 75 per cent for three channels, and 60 per cent for four channels.

For theatrical use the amplifiers should be capable of reproducing all frequencies within the range of 50 cycles to 10,000 cycles.

PUBLIC-ADDRESS SYSTEMS

By R. J. Kowalski

Public-address systems fall in two general classifications according to application. A system for amplifying speech or music presented directly to a large audience, whether in an auditorium or out in the open, is appropriately named a sound-reinforcing system. A system by which a speaker at a central location addresses people simultaneously at various locations is known as a paging system. In general, the equipment specifications for sound reinforcing, particularly indoors, are more exacting than those for paging applications.

8. INDOOR SOUND-REINFORCING SYSTEMS

The essential requirements of a sound-reinforcing system are that it must pick up all desired sounds and project them, unaltered, with sufficient intensity and distribution to all listeners within a given area. For maximum effectiveness, the system should function so as not to detract the attention of the listeners from the performers.

When installing a reinforcing system in an auditorium, particular attention should be given to the acoustic conditions. A certain amount of the sound energy emanating from the loudspeakers finds its way back to the microphone. Unless the difference between the level of the sound energy leaving the speakers and that arriving at the microphone is greater than the gain of the system amplifier, the system will oscillate because of acoustic feedback. Directional microphones aid in avoiding feedback, because, with them, it is possible to position the axis of maximum response toward the desired sound and the null axis toward the reflected sound. Mounting the microphones in acoustically treated compartments in the footlight trough also shields them from considerable reflected sound. Unless feedback can be corrected by repositioning the microphones or redirecting the loudspeakers it will be necessary to limit the gain of the amplifier and hence the amount of reinforcing obtained. However, if the auditorium is well designed acoustically, there should be no difficulty in getting sufficient amplification before feedback occurs.

The sound-reinforcing equipment in large auditoriums and theaters is usually custom built and permanently installed. The microphones and preamplifiers are broadcast quality, and the main amplifiers and loudspeakers are generally of the type used in sound-motion-picture reproduction. The amplifiers should have uniform frequency response over the complete range of audio frequencies from 50 to 10,000 cycles. The amplifier output capacity should be such that at least 1 acoustic watt per 1000 ft of floor space can be delivered. The amplifier should be able to deliver this amount of power with less than 2 per cent total harmonic distortion over the range from 50 cycles to 5000 cycles. The loudspeaker system should have directional horns to distribute the sound energy efficiently in the proper area. Many theaters employ two-way speaker systems for their reinforcing systems as well as for the movie-sound system. The two-way system assures uniform, highly efficient operation over the complete audio spectrum. The input equipment should be highly flexible to accommodate a wide variety of microphone combinations because different types of programs require different pick-up arrangements. See article 2. Where the pick-up area is great, such as an entire stage area for vaudeville and stage productions, microphones are usually located in the footlight troughs, suspended from the scenery drops, and sometimes concealed in stage props. The optimum spacing of footlight microphones has been found to be 8 to 10 feet. In order to control the input from all these microphones to get the proper balance in sound levels, a mixer is required with controls for each input. The most satisfactory location of the mixer console is somewhere in the audience area, preferably at the head of the balcony where the operator can hear the direct sound from the system speakers and adjust the various inputs for best balance.

An excellent example of a high-quality sound-reinforcing system is the one at Radio City Music Hall, New York City. Because of the width of the stage, a three-channel reinforcing system is used to obtain the best possible illusion. The stage microphones are split into three groups: right stage, center stage, and left stage. Each group is fed through its separate section of the mixer console, then to its separate amplifier, and thence to the speakers over the proscenium arch. The speakers, too, are split into three groups which are fed by the three amplifier channels. Therefore, a sound originating on the right side of the stage is picked up by the right-side microphones, amplified by the right channel amplifier, and reproduced by the right group of loudspeakers. As an actor moves across the stage the sound moves with him, creating an excellent illusion. Sound reproduction in this system is so well balanced that, despite the tremendous size of the auditorium, the people in the remotest corners can hear as well as those seated up front.

Besides these large, expensive systems, many small portable systems are available commercially that can be used for meetings, banquets, etc., in small rooms. These systems generally contain microphones, folding microphone stands, amplifier, and loudspeakers all in one compact carrying case. The sides of the case serve as baffles for the speakers when they are put in use.

9. OUTDOOR SOUND-REINFORCING SYSTEMS

Sound-reinforcing systems for outdoor use are generally higher powered than indoor systems, because the average area to be covered is greater, and the power per unit area is also greater since there is no beneficial reverberation to augment the direct sound. In order to utilize the available sound energy most efficiently, highly directional loudspeaker horns are used almost exclusively for outdoor work. Since the low-frequency cutoff on practical sized horns of this type is well up in the audio-frequency range, the system fidelity is not generally as good as that found in indoor systems. The fidelity could be improved with a more expensive speaker set-up; but, in most cases, intelligibility is more important in outdoor systems than fidelity. Since the low-frequency tones do not contribute materially to intelligibility, they may be sacrificed in the interest of holding down the practical size of the equipment. As in indoor systems, the loudspeakers should be located to give the best illusion; but they should be carefully directed to prevent any direct energy from reaching the microphone and causing feedback. This problem is obviously much less serious in outdoor than in indoor installations.

For indoor work, the ribbon-type velocity microphone is the most popular because of its uniform frequency response, its useful directional pattern, and its ability to pick up sounds at a considerable distance. In outdoor work this microphone is not too satisfactory because it is somewhat fragile and is subject to extraneous noises generated by the wind disturbing the ribbon. A much better microphone for this application is the dynamic pressure type.

10. PAGING SYSTEMS

Although the sound-reinforcing application of public-address equipment is probably more familiar to the layman, paging systems have become even more important. During the war years, large permanently installed plant broadcast systems became vital parts of most big factories, making it possible to locate key men quickly in the acres of floor area, and to coordinate production activities throughout the plant. Announce systems of many types, installed on practically all fighting ships of the fleet, proved to be the most important means of internal communication. New-type announcing systems for schools make it possible to communicate quickly with all rooms and to distribute educational programs as desired. Besides these, there are hundreds of more commonplace applications like train announcers at railroad stations and call systems in hospitals, hotels, restaurants, and other business establishments.

PLANT BROADCAST SYSTEMS. Though the leading manufacturers are building equipment components specifically designed for industrial use, there is no such thing as a universal industrial sound system. The desired services, plant layout, and conditions of operation vary so greatly that each installation becomes a custom-engineered job. Some of the many services furnished by a properly designed and managed industrial system are:

1. Paging. The telephone operator or a special paging operator, through announcements to selected areas of the plant, can quickly locate personnel.
2. Emergency and alarm. By using combinations of special signal generators and verbal announcements the system is highly effective in fire, damage, and accident control.
3. Time signals. The system may be connected to the main time clock to broadcast time signals at preselected intervals.
4. General announcements. The paging operator or one of the plant executives may use the system to supplant the bulletin boards for messages pertaining to plant operations.
5. Work music. To increase the efficiency of personnel, planned music programs may be given at periodic intervals during the work period. The source of the music may be recordings, a wired-in service, or radio programs.
6. Entertainment. During rest or lunch periods, programs may be presented by live talent, such as employee groups or visiting celebrities.
7. Morale building. The system may be used to bring about more personal contact in personnel relationship through inspirational messages, drives, safety campaigns, and announcements of general interest.

In recent years many scientific tests have been made on the value of music in industry. These have proved that periodic musical intervals have a beneficial effect on most types of workers, but especially on those engaged in repetitive manual operations associated with modern assembly-line manufacture. The chief effects are the relief of fatigue and boredom and the dispelling of nervous tension. Tangible results have been accomplished in the reduction of labor turnover, reduction of accidents, increased production, and improved quality of product.

The central control equipment of the average plant broadcasting system is located in a sound-treated room that serves as a studio. The central control console, or mixer, is located here to control the various inputs. These inputs may include a paging microphone, one or more studio microphones, one or more executive microphones in the executive offices, one or more radio inputs, one or two phonograph turntables, time clock signal generator, and possibly a fire or other emergency alarm signal generator. The mixer console should include an attenuator to control each input and a master attenuator to control the overall level of any combination of inputs. It should also include a volume-indicator meter to permit the operator to maintain the proper signal level on the system. The output of the mixer is fed into a program amplifier which in turn feeds a group of zone selector switches on the operator's console. In almost any setup it is best to divide the plant into convenient zones or areas for programming and paging. In a simple system the division might merely be "offices" and "factory," but in a larger plant an individual zone might be established for each building or each department. This makes possible selective announcements or paging calls to any one part of the plant. The output of the zone switches is fed to the individual zone power amplifiers that drive the loudspeakers in each zone. If the plant is small and all zones are in a single building, it is best to install all the zone power amplifiers in the studio or equipment room so that they will be in a central location to facilitate service work. However, if the plant is spread over considerable area, or in many buildings, it is best to mount the zone power amplifiers in the buildings they serve. The input signal can be fed from the studio at the zone power amplifier at zero level over standard telephone lines which normally link most buildings of a large plant. This is more economical on amplifier power than trying to feed high-level energy over long lines.

In a larger system it might be desirable to have provisions for sending two different programs to different areas simultaneously. This is especially helpful when an urgent paging announcement has to be sent to one zone during a regular musical program. The particular zone can be paged on a separate channel without disturbing the music going out to the other zones. To do this, it is necessary to add a second program amplifier and another volume indicator meter to monitor the level on this second channel. Each input switch should be a three-position switch so that, besides an "off" position, any of the inputs may be connected to either of the two program amplifiers. The zone selector switches should also be three-position switches so any zone may be connected to either of the two program channels or may be completely disconnected. In plants subject to serious emergencies, the switching arrangement may be so arranged that, whenever the alarm signal generator is sounded, it takes priority over all other inputs and is automatically fed to all loudspeakers regardless of the position of the zone switches. Alarm contactors to energize the signal generators may be located throughout the plant at critical points.

Associated with the studio equipment should be a monitor speaker with a switch which enables the system operator to monitor audibly any program going out to the plant or to check on proper tuning of the radio receiver before connecting it to the program bus. Adjacent to the studio should be a suitable storage room for storing the record library for musical programs as well as spare microphones, etc., for the system. Records for in-industrial use should be selected with some care. In general popular records are suitable for factory working periods with possibly light classical music for the plant restaurant during the lunch hour. Considerable research has been done on the psychological effects of different types of music. By reference to the bibliography following this article, additional information on this subject may be obtained. From a practical standpoint, since most factories are rather noisy the recorded music selected should have fairly constant level. If a recording having very loud and very soft passages is reproduced in a noisy area, the low passages will be lost completely unless the overall level is made so high that the loud passages are annoying.

The primary design consideration for equipment for industrial use is ruggedness. In many cases the equipment will have to operate continuously through a 24-hour working day. All components should be conservatively rated so they are used well within their capacity and hence will give long life. Precautions should be taken in the design of the switching system so that the equipment cannot be damaged by improper operation of the controls. Occasionally inexperienced personnel will attempt to operate the equipment, and so these safeguards are necessary.

The necessary overall fidelity of the equipment varies with the application and the noise levels encountered. If the system is to be used solely for verbal announcements, an overall frequency response of 300 to 3000 cycles is adequate; however, if music is to be reproduced, the response should extend from 50 to 10,000 cycles. This wide frequency range is not merely to give high-fidelity reproduction but also to add needed definition to music being reproduced under adverse conditions. Noise generally occurs at specific frequencies. If the music is reproduced with a limited band of frequencies, some of the frequencies will coincide with those of the noise and hence will be masked. The loss of certain notes reduces the definition of the music and makes it hard to follow. Of course, the difficulty can be overcome by making the volume louder to override the noise, but this might make the music annoying. By extending the frequency range, the definition may be improved without increasing the overall volume.

Plant noise levels and the shape of the building determine the size, type, number, and locations of loudspeakers. In high-noise areas, best results are obtained from horn-type loudspeakers which can direct the sound energy to the important areas. However, if the noise level is below 90 db, well-baffled cone-type loudspeakers should be used. It is possible to get more uniform distribution over wide areas with this type of speaker. The power required for any given zone may be computed fairly accurately by means of the formula given in article 5. Where the existing noise level is above 100 db, the power as computed by the formula should be increased accordingly. Most standard types of microphones are suitable for reproduction from a studio; but if the microphone is located in a high-noise-level area out in the plant, a specially designed close-talking microphone should be used to reduce the response to unwanted room noise. A phonograph turntable for industrial use should be weighted and dynamically balanced, and it should be driven by a heavy-duty motor to insure constant speed. Home-type record changers should never be used in an application such as this.

Before actually selecting the equipment for a particular plant, a thorough plant survey should be made. This survey should supply the following information:

1. Noise. All sources of noise should be located on a copy of the plant's floor plans as well as the average level of the noise in all areas. This information can be quickly obtained with a portable sound-level meter.

2. Coverage. On the basis of the above measurements, the types of speakers should be selected and their locations marked on the print.

3. Plant zoning. All areas should be zoned with reference to industrial operations for determining switching requirements.

4. Locations. The location of the studio, input sources, amplifier equipment, and controls should be determined and marked on the prints.

5. Special considerations. Any special information such as desired priority of signals and temporary microphone locations should be noted.

On the basis of the above information it will be possible to engineer a system that will exactly fill the needs of the plant. To fill the needs of a wide variety of different plants, some manufacturers have designed a series of very flexible sectionalized units which may be put together like building blocks to make up any desired type of system out of standard equipment.

NAVY ANNOUNCE EQUIPMENT. A modern fighting ship is a complex organization with a huge staff of personnel engaged in a wide variety of activities necessary to operate it. To coordinate all these activities it is necessary to use a shipwide, selective, announce system. When the ship is under way, the orders originate on the bridge; when the ship is in port, the ship's control center is the quarter deck, hence orders originate from this station. The ship is divided into sections so that orders may be issued to selected sections. In addition, each section is subdivided into smaller subgroups which can be separately disconnected in the event of battle damage; then if one speaker line is shorted, it may be cut loose so that it does not affect the rest of the system. Besides the shipwide general announce system, a large ship has many more specialized systems such as: (1) the engineer's announce system between the engineering log room and the various fire rooms, boiler rooms, and other engineering spaces; (2) the damage control system between the damage control office and several damage control stations; (3) turret control systems for communication between the turret captain in each of the primary battery turrets and the gun pointer, the gun layer, and the various shell-handling and power-handling decks below the turret; and (4) the secondary battery announce system communicating between the gun directors and the gun mounts of the secondary batteries. Moreover, there are high-powered systems using a single loudspeaker that can be directed for ship-to-ship, ship-to-dock, and ship-to-plane sound communication, and a variety of low-powered intercommunication systems between specialized points.

The requirements for ruggedness in ship equipment are even more exacting than in equipment for industrial uses. All ship equipment must be shockproof, and loudspeakers on the weather decks should be weatherproof, watertight, and blastproof to stand the pounding seas and the terrific concussion of gunfire. Microphones are generally of the close-talking dynamic type to reduce the response to undesired noises.

SCHOOL SYSTEMS. Small specialized versions of industrial systems have been designed for school use. The control console is located in the principal's office, where announcements and programs originate. The control console has a selector switch for each room so that announcements may go to any room or any group of rooms. A special connection with a talk-back amplifier makes it possible to use each classroom speaker as a microphone. In this way, it is possible to establish two-way communication between the principal and any of the teachers. Available inputs besides microphones include a radio tuner and a record player.

The main features of school equipment should be simplicity of controls and relatively low cost, which is achieved by a slight sacrifice in overall quality of reproduction, while maintaining a safety factor in the choice of component parts to insure dependability of operation.

BIBLIOGRAPHY

Burris-Meyer, H., Music in Industry, *Mechanical Engineering*, January 1943.
Cardinell, R. L., Principles of Programming Music in Industry, Amer. Soc. Composers, Authors, and Publishers (1944).
Halpin, D. D., Industrial Music and Morale, *J. Acoustical Soc. Am.*
Music in Industry, Industrial Recreation Association (1944).
Selvin, Ben, Programming Music for Industry, *J. Acoustical Soc. Am.*, October 1943.

SOUND RECORDING AND PROJECTION

11. RECORDING PRACTICES

By O. B. Gunby

Sound motion pictures are released extensively, either 16 mm or 35 mm, depending upon the application. The use of 16-mm film is on the increase in advertising and educational fields. The majority of the studios make all their pictures on 35-mm film, even though some of them will be reduced to 16 mm before release, because of the greater flexibility obtainable from the 35-mm equipment commercially available.

Recent trends in the design and operation of sound-motion-picture equipment include the general use of electronic mixers (volume compressors) or limiters to vary the volume range of the recordings or to prevent overloads on loud signals. Also, particularly in the recording of music, synchronously driven acetate recordings are often made so that an immediate playback of the recorded material is possible, thus permitting a quick and accurate check on quality.

A block diagram of a typical production-type dialogue channel is shown in Fig. 1. The same type of a channel may be used for recording sound effects or for the recording of

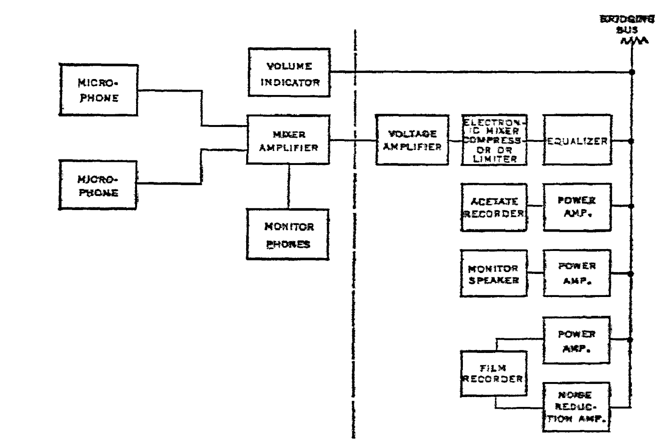


Fig. 1. Simplified Scoring or Dialogue Recording Channel

music, which is usually referred to as scoring. For scoring, however, it is customary to use more microphone inputs than are indicated for a dialogue channel. Since the picture is usually made before the music is recorded, a motion-picture screen is generally provided on a scoring stage so that the orchestra leader can watch the picture while he is directing the orchestra.

In a dialogue channel the microphones are usually located 4 ft or less away from the actors, but out of range of the camera. The microphones are usually mounted on long adjustable booms so that they can be moved to follow the action closely and yet avoid the camera. Microphones of the unidirectional type are used frequently because of their ability to discriminate between the wanted sound from the set and the extraneous noises from other directions.

The sound stage in which the recording channel is used must be constructed to exclude external noises. This frequently involves double wall construction with intervening air spaces and accoustic treatment on the interior to provide the desired reverberation time.

The mixer is located near the set in a position where the operator has an unobstructed view of the action being recorded. This helps him to anticipate changes in mixer adjustment to suit the sound source. Monitoring is accomplished with high-fidelity headphones. The remainder of the recording channel may be located in a small room on the sound stage, in a recording truck parked adjacent to the stage, or in a centrally located building on the studio grounds and connected to the sound stage by suitable transmission lines.

Figure 2 is a block diagram of a typical rerecording channel. It is essential that the console be installed in a room having acoustic properties comparable to those of an average

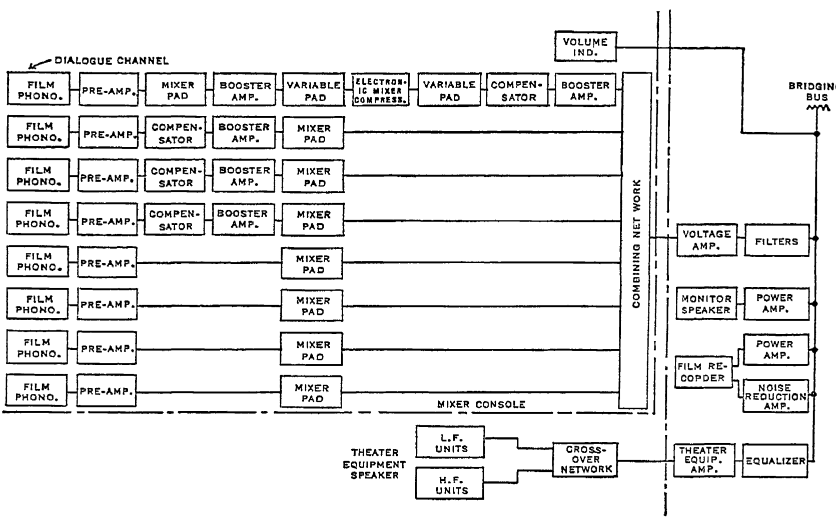


FIG. 2. Simplified Rerecording Channel

theater so that the desired sound quality may be determined through the monitor speaker system as the numerous sources are blended together to make a final sound track.

The various sound sources are usually recorded on separate films in order to provide the desired degree of flexibility. In general, the sound tracks will include dialogue, music, and any number of sound-effect tracks that may be required. These films are threaded in separate film phonographs whose audio outputs are individually controlled by the operator at the mixer console. The threading of the films is indicated by suitable marks so that sound and picture are synchronized. The various film-handling machines are driven by Selsyn motors which provide an electrical interlock during the operating cycle.

During a rehearsal the film phonographs and the projector are run by the Selsyn driving system, and the operator varies the audio signals from the film phonographs to fit the mood of the picture. Several rehearsals are usually required in obtaining the desired effect. Once this has been achieved, the films are rethreaded and another run is made, including the film recording machine. During this run the operator endeavors to repeat the changes in level, etc., that were made during the successful rehearsal. The procedure is repeated, if necessary, until a satisfactory rerecording is made.

During the rerecording process, in addition to changes in level that are made in the incoming signals, changes may also be made in the frequency response to adjust for day-to-day differences in the original recording or to obtain certain special effects such as telephone quality, old phonograph quality, and so forth.

The more progressive studios usually have facilities for the control of reverberation. A reverberation chamber is one means of control. A portion of the sound requiring reverberation is fed into a loudspeaker in the chamber; the sound is picked up by a microphone located in the same room and mixed with the original sound to obtain the desired effect.

A third type of recording channel, the single-film system, is shown in Fig. 3. This channel is generally used in the field for the original recording of newsreels and to a lesser extent for recordings on locations impractical for recording with the more cumbersome production-type channels. As the name implies, this system uses one film on which both the picture is photographed and the sound is recorded.

The more modern equipments use a specially designed camera having in it a mechanically filtered drum on which the light beam from a compact recording optical system can be focused. With some equipments, a conventional camera is used and the sound is recorded directly on the camera sprocket.

The amplifier is designed to have small size, light weight, and low power drain, and the facilities provided are kept to a minimum. Power for the amplifier, exposure lamp, and camera motor are often provided by a single small low-voltage storage battery.

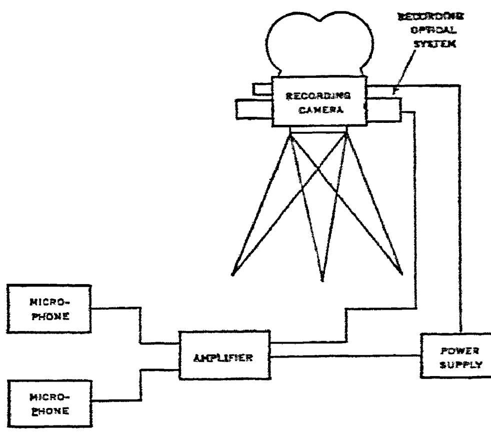


FIG. 3. Single Film Recording Channel

Some newsreel channels use class B pushpull recording since this type of track provides excellent noise reduction by optical means and does not add to the size, weight, or power drain of the equipment. This track requires rerecording to standard track before it can be released to theaters, but since it is always rerecorded to provide sound effects, music, etc., this requirement presents no problem in the production of the newsreels.

Modern recording equipments are capable of producing films having a flat frequency response over a range greater than 30 to 10,000 cycles and a volume range in excess of 50 db. However, in practice it is usually found desirable to restrict the frequency range from about 70 to 7000 cycles and the volume range to approximately 20 to 25 db. This limitation in frequency and volume range is a compromise resulting from the necessity of evolving a technique that provides commercially acceptable sound quality while recognizing the following variable factors: (*a*) reasonable quality tolerances for each of the steps in sound motion picture production, (*b*) the wide variety of conditions under which sound films are reproduced in the many theaters.

12. PROJECTION PRACTICES

By J. D. Phyfe

The sound-reproducing system of a modern motion-picture theater is the combined product of many highly specialized arts and sciences, embracing the fields of optics, acoustics, electronics, and mechanics.

The following description of the major component items of a typical sound-reproducing system will illustrate how these components are combined into a complete system.

The equipment housed in the projection room usually consists of two picture projectors and associated lamp houses, two soundheads which are mounted below the projector mechanisms, and an amplifier system. The projectors and soundheads for all theaters are quite similar, the power output rating of the amplifier and loudspeaker systems being modified to compensate for changes in the seating capacities of various theaters.

The sound-reproducing industry has set progressively higher standards of performance, as exemplified in constant development work fostered by all manufacturers of theater equipment. Tentative standards of reproduction have been established by the Research Council of the Academy of Motion Picture Arts and Sciences.

The film, upon which are photographed both the picture and sound records, must be moved through the mechanism of the soundhead at a constant rate of speed to insure a high quality of sound reproduction free from "wows" or "flutter." These terms denote minute variations in the speed of the film.

Advances have been made in the design of the film-moving mechanism to reduce to a minimum irregularities in speed of the film. Two types of film motion filter are in current use, both employing a film-driven rotating drum coupled to a flywheel. One type,

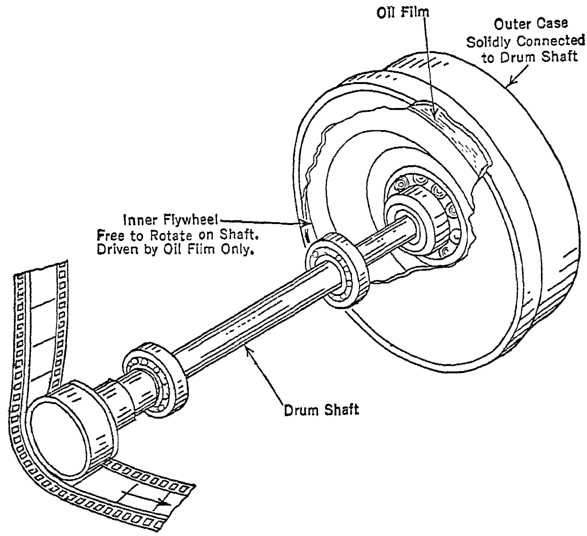


FIG. 4. Soundhead Film Motion Filter (Rotary Stabilizer)

known as the Rotary Stabilizer (see Fig. 4), consists of two flywheels connected by a viscous medium. The other type utilizes a solid flywheel in conjunction with a dampened idler roller for its filtering action.

A picture of a modern soundhead is shown in Fig. 5. The view is of the "operating side" of the unit through which the film passes.

Light from a source termed an "exciter lamp" passes through an optical system where the dimensions of the beam are rigidly defined into a narrow slit 0.00125 in. wide and 0.084 in. long. This slit image is focused upon the sound track of the film, which is moving

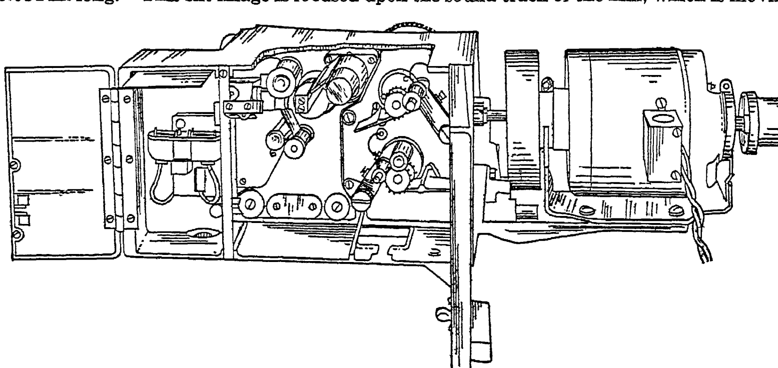


FIG. 5. Operating Side of Modern Soundhead

and presenting a continuously variable ratio of clear film to the dark or exposed area of the film that is being scanned by the light (Fig. 6). The film effectively serves to control the transmission of the light in conformity with the light and dark areas that comprise the sound record. The variation in the transmission of light is translated into a corresponding variation in current by means of a photoelectric cell. (See Section 15, article 7.)

Because the photocell currents are weak, it is necessary to provide a means of amplification; this may consist of a one- or two-stage voltage amplifier. Amplifiers are usually separate units, either incorporated into the soundhead or mounted on the front wall of the projection booth near the soundhead. They may also be placed in the main amplifier rack and coupled to the soundhead by means of transformers or suitable low-capacity coaxial cables.

Further amplification of the photocell voltage is furnished by the main or power amplifier, raising the low-level currents to a satisfactory value where they can be made to operate the theater loudspeaker system.

In order that the projectionist be informed constantly of both the volume level and quality of the sound, a small monitor loudspeaker is installed in the projection room.

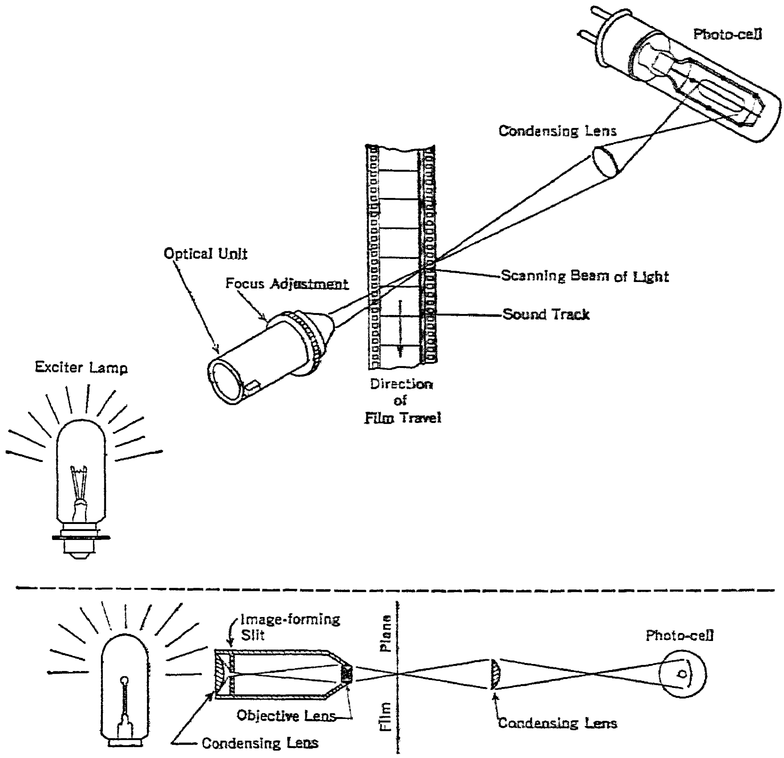


FIG. 6. Optical System of Modern Soundhead

Equalizers are frequently employed to adjust the electrical response characteristics of the amplifier system so as to provide optimum acoustical results in a given theater.

It is common practice to select the output of the desired soundhead by alternate switching of the exciter lamp currents or by selecting the audio output of either soundhead by means of a "change-over switch" or "fader."

The projectionist is notified of the proper time for making the change-over between projection equipments by two small cue marks. These marks will appear in the picture area of the film and are visible on the screen. The cues are placed several seconds apart on the film as it passes through the outgoing projector. The first mark signals the operator to start the motor of the incoming machine. The second cue mark, appearing shortly after the machine has attained full operating speed, marks the point of actual change-over. The picture is switched by an electric dowser actuated by a foot switch operating in synchronization with the sound change-over.

TWO-WAY LOUDSPEAKER SYSTEM. The loudspeakers employed in theater sound reproduction are located behind the picture screen. Small perforations in the screen, not noticeable from the seating area of the theater, permit the sound to pass readily through the screen. A typical two-way loudspeaker system is shown in Fig. 7.

The wide frequency range and the large power-handling requirement of modern theater loudspeaker systems cannot be met by a single speaker mechanism. The result has been the development of the two-way loudspeaker in almost universal use today. For the higher frequencies, a unit having a small light-weight diaphragm coupled to a multicellular

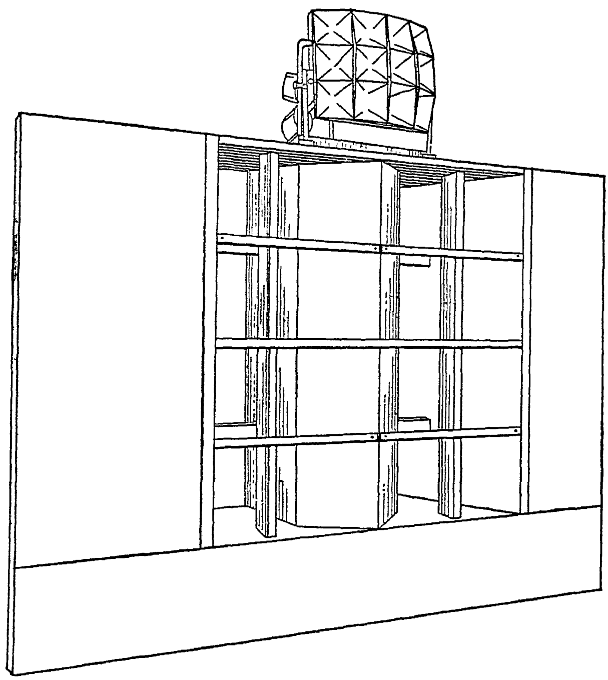


FIG. 7. Two-way Theater Loudspeaker System

directional horn is employed. The low-frequency portion of the signal is assigned to a unit having a larger diaphragm coupled to a very large horn or baffle. The division of the high- and low-frequency components is accomplished electrically through a "cross-over network." The schematic diagram of such a network appears in Fig. 8. The cross-over frequency is in the region of 400 cycles.

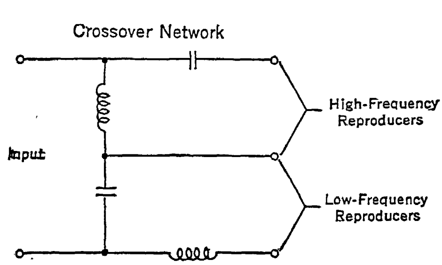


FIG. 8. Schematic Diagram of Crossover Network for Two-way Loudspeaker System

A block schematic of a complete theater sound-reproducing system is shown in Fig. 9. The relative circuit positions of the components covered above may be readily observed.

RECENT DEVELOPMENTS. Recent achievements in motion-picture-sound engineering are the development of the control-track system of reproduction and the drive-in type of theater. The first-named system employs a control track consisting of variations in the area of exposure of the small portion of film lying between adjacent sprocket holes.

The system has been utilized to produce a 96-cycle tone, the frequency being governed by the number of sprocket holes per second passing a small scanning light. The variations in light actuate a separate photoelectric cell as described earlier in this chapter. The photocell current, after amplification, is rectified and used as a control voltage to regulate the output level of the main theater amplifier and to cut into operation an auxiliary loudspeaker system placed at each side of the picture screen. This complete system permits a tremendous volume range not otherwise obtainable and provides a wider source of sound

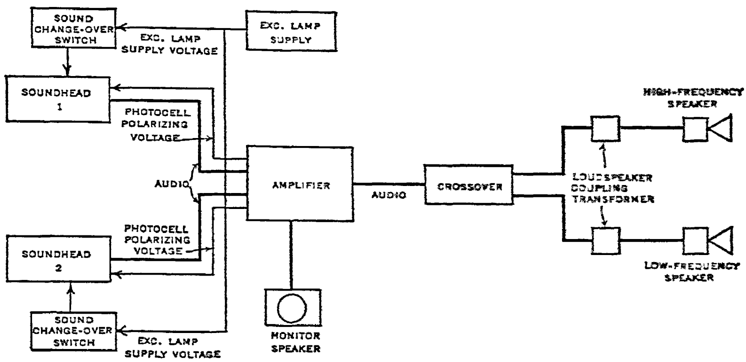


FIG. 9. Overall Block Diagram of Theater Sound-reproducing System

during loud passages than speakers alone placed behind the screen; it heightens the dramatic effect of certain loud passages.

DRIVE-IN THEATERS. The popularity of the drive-in type of theater has increased considerably during the last few years. The picture is shown on a large outdoor screen. The patrons remain seated in their automobiles, which are located in an arc to permit viewing the picture through the windows. The trend is toward the use of individual loudspeakers which may be placed inside the cars.

BIBLIOGRAPHY

Kimball, H. R., Applications of Electrical Networks, *J. Soc. Motion Picture Engineers*, Vol. 31 (October 1938).

Levinson, Nat., and L. T. Goldsmith, Vitasound, *J. Soc. Motion Picture Engineers*, Vol. 36 (August 1941).

Reiskind, H. I., Reproducing Systems, *J. Soc. Motion Picture Engineers*, Vol. 36 (August 1941).

RADIO TELEPHONE BROADCASTING

By Howard A. Chinn

Radio broadcasting is a means for delivering intelligence for general reception at distant points. A complete system consists of: (*a*) a radio broadcasting transmitting system; (*b*) the medium through which transmission takes place; (*c*) a number of receiving installations.

A radio broadcasting transmitting system consists, essentially, of:

1. A studio, stage, theater, auditorium, or other suitable place for the performance that is to be broadcast. (See article 1, above.)
2. An acoustoelectric device (microphone) actuated by sound energy and delivering electrical energy. (See article 2, above.)
3. Amplifiers for increasing the amplitude of this electrical energy. (See article 3 above.)
4. Control equipment for the regulation and adjustment of this electrical energy. (See articles 3 and 4 above.)
5. Wire lines to carry the electrical replica of the original sound waves from the studio to the radio transmitter.
6. Radio transmitter for converting this electrical energy into radio-frequency energy.
7. Antenna system for radiating the radio-frequency energy into space.

A schematic diagram of a typical broadcast transmitting system, showing the general type of circuit layout employed, is given in Fig. 1. A single studio and a single remote pick-up point are represented, each such point requiring a duplicate of the equipment

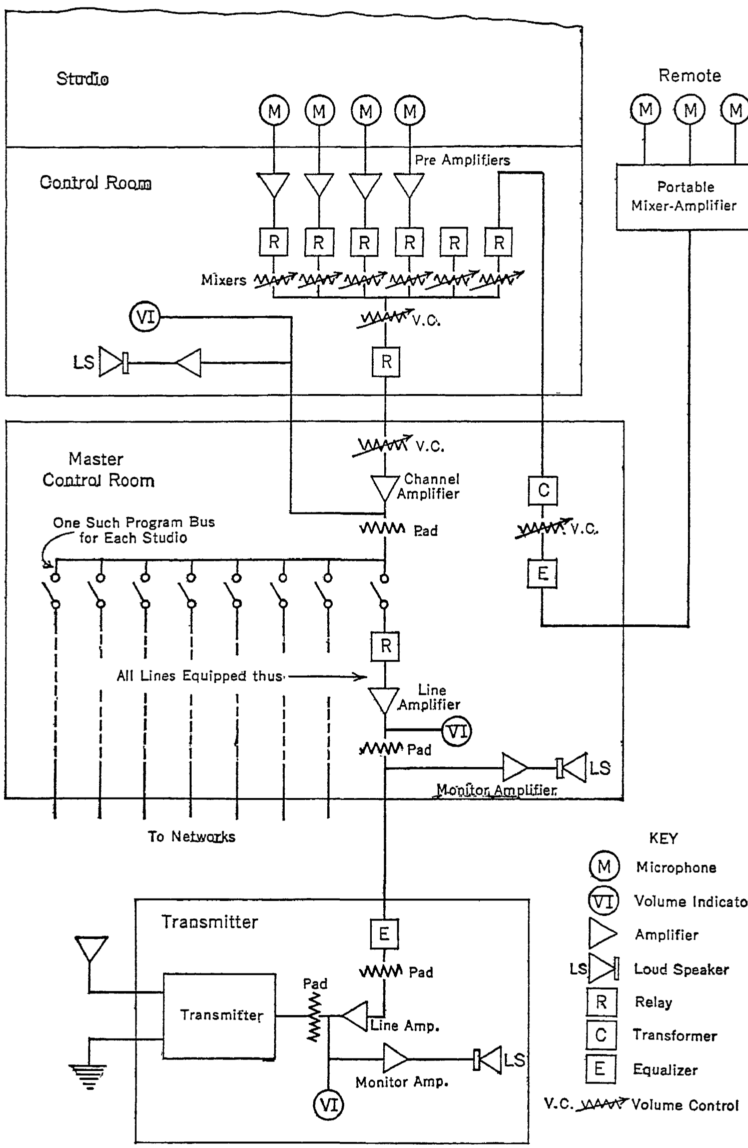


FIG. 1. Broadcasting System Layout

shown, up to and including the relays associated with the outgoing lines to the various networks.

The first four items listed above as parts of a complete broadcasting system have been described in detail in the opening articles of this section, "Audio Facilities for Sound Systems." The remaining items listed are more or less peculiar to broadcasting systems and are covered below.

13. PROGRAM DISTRIBUTION SYSTEMS

NETWORK SWITCHING EQUIPMENT. The key stations of a network of broadcasting stations must provide means whereby the output of any studio may be distributed to any of the networks or combination of networks radiating out from the city in which the key station is located. Frequently different programs, coming from different origination points, are simultaneously sent to the various legs of the network radiating from the key station. In order to accomplish these operations switching means are used which permit the connection of a line amplifier across the program bus of the desired program source.

The facilities for switching are usually such that the proper studio and network line-up may be arranged previous to "air time" but without actually connecting the studios involved to their respective networks until a master switch is operated. Upon the proper cue, or at the proper time, operation of the master switch connects the various studios involved to the right networks. The actual switching is seldom accomplished by manually operating the switches but rather through the medium of conveniently located relays which are remotely controlled from the operating desk.

BRIDGING AMPLIFIER. The purpose of the bridging amplifier is to isolate the outgoing "radio" lines from one another and to provide a means of connecting any number of outgoing lines to any program source, at will, without causing any unbalancing or impedance mismatch of the equipment line-up. If two or more outgoing radio lines feeding different networks, but carrying the same program, were to be connected in parallel and thence to the output of the program amplifier, then, should any noise, ground, or other fault develop on one line, it would affect the operation of the others. By placing a bridging amplifier (which is, of course, a one-way device) in each line, complete isolation is effected and there is no possibility that one line will affect others being fed from the same studio.

The bridging amplifier also permits the connection of any reasonable number of lines to the output of a given studio amplifier without causing an impedance mismatch which would adversely affect the operation of the system. To accomplish this connection, the output of the program amplifier is terminated in a resistance of the proper size, thereby presenting a practically constant load for the amplifier. The input impedance of the bridging amplifier is then made very high and is "bridged" across the desired program bus without appreciably affecting the load impedance being presented to the output of the program amplifier.

A bridging amplifier is associated with each of the outgoing lines leading to a local transmitter or to a network of stations.

BIBLIOGRAPHY

Chinn, H. A., CBS Hollywood Studios, *Proc. I.R.E.*, July 1939, p. 421.
Rackey, C. A., Network Broadcasting, *Elec. Eng.*, January 1941, p. 16.

14. PROGRAM LINES

Telephone lines are employed for the purpose of transmitting a program from one studio or station to another station. The facilities involved may be divided into two classes: (1) local lines for connecting the studio remote pick-up points, such as athletic fields, theaters, and hotels, and also those lines used for connecting the studios to the local transmitter; (2) long lines interconnecting a network of transmitting stations throughout the country.

LOCAL LINES. When the stations involved are in the same city the line connecting facilities are known as loops. These relatively short lines may, by the employment of proper terminal equipment, be made to have an essentially flat transmission vs. frequency characteristic over the entire range of audio frequencies necessary for high-fidelity broadcast service (see Section 16, article 18). In order to obtain this desirable feature the natural attenuation characteristics of the lines, which for the most part are cable circuits, are modified at the receiving terminals by means of an attenuation equalizer.

LONG LINES. If the stations to be interconnected are in different cities the connecting facilities consist of special telephone lines which are either non-loaded open wire or loaded cable circuits (see Section 17, article 18). Present cable facilities are loaded at intervals slightly in excess of $^1/_2$ mile. Amplifiers, equalizers, and phase correctors are installed at approximately 50-mile intervals on cable circuits and about 125 miles apart on open wire facilities. The cable circuits have automatic regulators installed about every 150 miles in order to keep the loss of the circuit independent of the temperature along the circuit.

Attenuation equalizers are employed on these circuits just as in local lines. Velocity correctors (see Section 5, article 10, and Section 17, article 18) are also necessary to compensate for the natural characteristic of the lines which results in an unequal time delay in the transmission of the various component waves of different frequencies. In circuits less than 500 miles long these devices would not be necessary, but with present circuit requirements of 2000 to 3000 miles they are indispensable.

Long-line facilities are available with overall transmission vs. frequency characteristics that are essentially uniform over the entire audio-frequency range necessary for high-quality broadcasting.

ATTENUATION EQUALIZER. (See also Section 17, Article 18.) An attenuation equalizer is bridged across the receiving terminals of a line in order to modify the natural characteristics of the line so as to provide a circuit having an essentially uniform transmission vs. frequency characteristic over the range of frequencies desired. The device is connected at the receiving end of the line in order to obtain the best ratio of signal-to-noise and interference on the circuit.

An attenuation equalizer is an electrical network which introduces a loss at each frequency such that the sum of the line and equalizer losses is the same for all frequencies over the useful range.

In its most elementary form the equalizer consists of a simple resonant circuit in series with a variable resistance. The frequency of the resonant circuit is selected so that, with the proper value of series resistance, the overall transmission characteristic of the circuit is as uniform as practical.

BIBLIOGRAPHY

Bode, H. W., Variable Equalizers, *Bell Sys. Tech. J.*, April 1938, p. 229.

Clark, A. B., and Green, C. W., Long Distance Cable Circuit for Program Transmission, *Bell Sys, Tech. J.*, July 1930, p. 567.

Cowan, F. A., Telephone Circuits for Program Transmission, *Trans. Am. Inst. of Elec. Eng.*, July 1929, p. 1045.

Cowan, McCurdy, and Lattimer, Engineering Requirements for Program Transmission Circuits. *Bell Sys. Tech. J.*, April 1941, p. 235.

15. BROADCASTING TRANSMITTER PLANT

A broadcasting transmitting plant consists of audio input equipment, modulator, radio-frequency generator and amplifier, radio-frequency transmission line, antenna tuning equipment, and antenna system.

The audio equipment associated with a transmitter plant provides facilities for such switching operations as are required, microphone and turntable equipment for local program origination in an emergency, and amplifiers for increasing the volume level of the program material received from the line connecting the studios to the transmitter. After being sufficiently amplified the incoming program material passes to the modulator tube which modulates the radio-frequency energy generated and amplified by the equipment supplied for that purpose (see Section 7, article 17; also Section 8, article 4). The resultant modulated radio-frequency energy may either be further amplified or sent directly to the antenna tuning equipment. The antenna is usually located a relatively short distance from the building housing the transmitter, and a radio-frequency transmission line is used to convey the energy from the transmitter to the antenna. The antenna tuning equipment is usually located at the base of the antenna in an appropriate protective shelter.

STANDARD A-M BROADCASTING TRANSMITTING ANTENNAS. (See Section 6, article 31.) For standard broadcasting (amplitude modulation in the 540-1600 kc band), the vertical-radiator antenna is generally used. The use of an antenna having an electrical height slightly in excess of 0.5 wavelength, and operated below the fundamental, results in the largest field intensities on the horizon for a given radiated power. At the optimum point of operation the electric field at the receiver resulting from the ground wave radiated by the antenna may be as much as 40 per cent greater than that obtained from a 0.25-wave antenna radiating the same power. This improvement results from the fact that more energy is radiated along the ground, where it is desired, and less up in the air. It does not follow from this, however, that the maximum coverage is secured by an antenna having this optimum height.

For the low-powered transmitter where the primary range is limited by the field intensity falling below the prevailing interference level, an antenna of the optimum height would probably result in an increased service area. The cost of such a radiator in comparison with the cost of increasing the power of the transmitter sometimes precludes its use, however.

For the high-powered transmitter the primary range is generally limited to that distance where "mushing" results from the admixture of the ground and the sky wave, at this point the strength of the waves being of about the same magnitude. In this case variations in the sky wave brought about by varying the height of the antenna are far more important in determining the primary range of the station than attendant variations in the ground wave. The height that is the best operating point for the greatest ground-wave intensity is not always the best height from the viewpoint of pushing out the incipient fading distance by the reduction of the sky wave. Hence, the best operating condition for maximum primary coverage is not necessarily that height which results in the maximum ground wave. The best height for a given antenna depends upon the attenuation of the ground wave, which in turn depends upon the effective conductivity of the soil, its dielectric constant, and the frequency of operation. In any event the optimum electrical height is likely to be between 0.5 and 0.6 wavelength.

The economical advantage of an antenna of this height depends upon the transmitted power. The initial investment and the cost of operation of the transmitting plant increase with the power, whereas the cost of the radiating structure remains practically constant. At the higher powers this type of radiator represents a good balance between the two investments.

DIRECTIONAL ANTENNAS FOR STANDARD (A-M) BROADCASTING. The application of antenna systems having definite directional properties to broadcasting purposes has been undertaken in a number of instances. Among the circumstances which have led to the installation of a directional system are: the need for suppressing radiation in a particular direction or directions in order to prevent interference with a distant station or stations operating on the same channel; the desire to suppress radiation in a given direction where no audience exists and to reinforce transmission towards the populated area, as for instance in a station located on a seacoast or to one side of a town which constituted its principal audience.

A suitable number of vertical antenna elements properly phased and spaced are usually employed in order to obtain the desired directional characteristic. By the proper combination of these antenna elements and their proper phasing almost any desired directional pattern may be obtained (see Section 6, article 29). Either vertical-wire antennas or towers insulated at their base are used for the antenna elements.

F-M BROADCASTING TRANSMITTER ANTENNAS. For f-m broadcasting (frequency modulation in the 88- to 108-Mc band), a horizontally polarized antenna system is employed. In general the antenna is non-directional in a horizontal plane. However, since most receiving sites are located within a few degrees of the horizon, it is advantageous to utilize an antenna system which directs the radiation towards the horizon. The gain realized by this practice permits the use of lower actual transmitter power for a given "effective" radiated power (effective radiated power is the actual power multiplied by the power gain of the antenna in the direction of the horizon). In practice, the cost of transmitters of various power levels must be balanced against the cost of directional antennas of various gains in determining the optimum combination.

The f-m broadcasting antenna must be located at a point of high elevation in order to reduce to a minimum the shadow effect on propagation of hills and buildings. To provide the best service to an area, a high antenna is usually preferable to a lower one with increased transmitter power.

STANDARD (A-M) BROADCAST STATION TRANSMITTER SITES. The selection of a good site for a standard (a-m) broadcast transmitter is a very complex problem which involves many considerations. The site should be selected with the view to providing:

1. Satisfactory coverage of the area comprising the population it is desired to serve. Usually this consideration will fix the maximum distance from the center of the city that the transmitter can be located.

2. Maximum coverage of adjacent populated areas consistent with fulfilling the above requirement.

3. Minimum population in the area immediately adjacent to the transmitter where the signal is likely to be so strong that special precautions may have to be taken to insure good reception from other stations.

4. Good soil conditions at the transmitter site. The conductivity of the soil within several wavelengths of the antenna has considerable bearing upon the efficiency of the antenna and the nature of its radiation characteristics.

5. Good power and program circuit facilities. If possible, two sources of power coming from different directions should be obtained. In order to obtain better regulation, it is often advisable to obtain power from a high-voltage line and have a local substation installed. Because of their relative immunity from storms, telephone lines in cable should be obtained.

6. Low cost of land. The size of the plot necessary will depend upon the size of the ground system, the spacing of the towers, and the distance between the anchors for the guys.

7. Good publicity value and accessibility. These are good assets for a station but may be overemphasized. Of course, roads leading to the transmitter should be usable in any kind of weather.

8. Immunity from floods, storms, sleet, etc., whenever possible, and ground suitable for good tower foundations. In some instances severe storms are localized in certain areas that can be avoided. Severe storms may cripple power and telephone facilities.

9. Proper location with respect to airports and airways.

10. Proper location with respect to large metal obstructions, buildings, etc.

F-M BROADCAST STATION TRANSMITTER SITES. The selection of a site for a f-m broadcasting station entails considerations somewhat different than those for standard broadcast stations. Many of the differences stem from the quasi-optical nature of the very-high-frequency-wave propagation. The transmitter site should be chosen with these factors in mind:

1. The location should be as near the center of the proposed service area as possible consistent with the availability of a site with sufficient elevation to provide service throughout the area.

2. The location should provide line-of-sight over the principal city or cities to be served. No major obstructions should be in the path.

3. The site should be so situated that the field intensity in the urban area is sufficiently great to provide satisfactory service in spite of the generally higher electrical interference in such areas.

4. Good power and program circuit facilities are required.

5. If the cite is a high building, consideration must be given to the problems of installing the antenna and the transmitter.

6. Cognizance must be taken of the possible hazard of the antenna to aviation.

BIBLIOGRAPHY

Brown, G. H., Directional Antennas, *Proc. I.R.E.*, January 1937, p. 78.
Chamberlain, A. B., and Lodge, W. B., The Broadcast Antenna, *Proc. I.R.E.*, January 1936, p. 11.
Lodge, W. B., The Selection of a Radio-broadcast Transmitter Site, *Proc. I.R.E.*, October 1939, p. 621.

16. BROADCAST FREQUENCY ALLOCATION

The frequency spectrum now known to the radio art extends over a wide range and includes frequencies having widely different characteristics. This spectrum is occupied not only by broadcasting services but also by other kinds of radio services such as communication with ships and aircraft, police services, and amateur, experimental, transoceanic, and transcontinental point-to-point communication, both telegraph and telephone. Exploratory work is still going on in the higher frequencies at the upper end of the spectrum and is directed in part to determining their usefulness for broadcasting purposes.

The nations of the world have agreed to devote certain portions of the radio-frequency spectrum to broadcasting purposes. The standard a-m (amplitude modulation) broadcast band extends from 540 to 1600 kc per sec and is used generally throughout the world. A band extending from 160 to 265 kc is used in Europe but not in this country for a-m broadcasting. Several narrow bands in the high-frequency spectrum (above 6000 kc) are also in use for long-distance a-m broadcasting services. Finally, a band extending from 88 to 108 Mc is used in this country for f-m (frequency modulation) broadcasting. The wave propagation characteristics of transmissions made in these various bands differ radically (see Section 10, article 24).

STANDARD BROADCASTING. The term "standard broadcasting" is applied to a-m stations operating in the band of frequencies from 540 to 1600 kc. Each station is assigned a particular carrier frequency. On the North American continent the assignable frequencies extend throughout the range in 10-kc intervals. Thus the assignments are 540, 550, 560, etc., up to 1600 kc, making a total of 107 distinct channels.

There are three classes of standard broadcast channels: clear, regional, and local.

A clear channel is one on which the dominant station or stations render service over wide areas and which are cleared of objectional interference within their ground-wave service areas and over all or a substantial portion of their sky-wave service areas.

A regional channel is one on which several stations may operate with powers not in excess of 5 kw. The ground-wave service area of a station operating on any such channel may be limited, as a consequence of interference, to a given field-intensity contour.

A local channel is one on which several stations may operate with powers not in excess of 250 watts. The ground-wave service area of a station operating on any such channel may be limited, as a consequence of interference, to a given field-intensity contour.

By assigning adjacent channels in widely separated areas of the country potential interference is minimized. In any one area it is common practice to separate the channels by approximately 30 kc. This leaves sufficient frequency separation to enable receiving sets to select one channel to the exclusion of all others in that area.

HIGH-FREQUENCY BROADCASTING. By international agreement high-frequency bands have been allocated for broadcasting services in the vicinity of 6, 9, 11, 15, 17, and 21 Mc. Transmissions in these bands are utilized for a purpose and in a manner entirely different from those in the a-m or f-m broadcast bands. High-frequency transmissions are primarily intended for long-distance broadcasts to distant colonial possessions, isolated territories, and overseas broadcasting. This type of service depends entirely upon the sky wave for reception as contrasted to regular broadcast transmissions which utilize the ground wave for primary coverage (see below). High-frequency transmission to distant points is not very satisfactory when reception is obtained with the relatively simple equipment available for the broadcast listener. Magnetic disturbances and atmospheric conditions seriously affect high-frequency transmissions and cause amplitude and selective fading and associated deterioration of tonal quality.

F-M (FREQUENCY MODULATION) BROADCASTING. Frequencies above 30,000 kc are referred to as very high frequencies. These waves are sometimes known as quasi-optical waves because their transmission characteristics resemble, in many respects, those of visible light waves (see Section 10, article 20). As a consequence the service range of a very-high-frequency broadcasting station, even if located on a high point so that the waves travel to the receiving station with a minimum of obstacles in their path, is limited to several tens of miles.

As compared with standard a-m broadcasting frequencies, very high frequencies present several advantages. Interference caused by natural atmospheric disturbances (static) is essentially non-existent, and therefore reception is markedly less dependent on seasonal influences. The service range of the station is more clearly defined and independent of any normal Heaviside layer conditions. The area over which a very-high-frequency station creates interference with other stations on the same or adjacent frequencies is not so great, compared to the useful service area, as in standard broadcasting frequencies. A substantial advantage exists in this respect that is of real assistance in very-high-frequency allocation. The dimensions of the receiving antenna can be small, and directional transmission and reception are relatively easy.

The disadvantages of very-high-frequency waves are inherent in their very nature. The high absorption during propagation limits the service range so that the covering of a large geographical area by this means on an economical basis presents a problem. Because of the quasi-optical character of the very-high-frequency waves there may be propagation shadows and areas of relatively poor reception, particularly near hilly terrain or high buildings.

In this country the band of frequencies from 88 to 108 Mc has been assigned for very-high-frequency broadcasting. The assignable frequencies extend throughout the range in 200-kc intervals. Thus the assignments are 88.1, 88.3, 88.5, etc., up to 107.9 Mc, making a total of 100 distinct channels. Frequency modulation, with a carrier swing of ± 75 kc, is used. The term f-m broadcasting is applied to this class of service.

Currently in this country there are two classes of f-m stations. Those designated as Class A are designed to render service primarily to a community or to a city or town other than the principal city of the area and the surrounding rural area. Class B stations are designed to render service primarily to metropolitan districts or principal cities and surrounding rural area, or to rural areas removed from large centers of population.

BIBLIOGRAPHY

Federal Communications Commission, *Standards of Good Engineering Practices concerning Standard and FM Broadcast Stations. Federal Communications Commission Rules and Regulations.*

17. BROADCASTING STATION SERVICE

STANDARD BROADCAST COVERAGE. In considering the probable service area of a standard broadcast station it is necessary to take into account the effects of both the ground and sky waves which are radiated by the transmitting antenna.

The ground wave (or direct ray) which travels directly over the surface of the earth from the transmitter to the receiver is unaffected in its propagation by meteorological or seasonal conditions and is of the same intensity during both the day and the night. The sky wave (or indirect ray) which traverses the Heaviside layer is subject to a great deal of variation in strength and character before reaching the receiving point (see Section 10, article 24).

The radiated energy which follows close to the earth, called the ground wave, is characterized by: (*a*) high field intensities near the transmitter; (*b*) attenuation to low values within a few tens of miles, depending upon the character of the ground, the power and frequency of the transmitter signal, and the type of transmitting antenna; (*c*) relatively steady values.

The energy which is reflected back from the ionosphere (chiefly evident after sunset), called the sky wave, is characterized by: (*a*) considerable field intensity at distances of hundreds of miles from the transmitter; (*b*) wide variation in field intensity from moment to moment, from night to night, and from year to year; (*c*) considerable variation with latitude of the transmission path and the earth characteristics in the vicinity of the transmitter, and some variation with frequency.

Because of its steady nature and the strong signals obtainable in areas near the transmitter, the preferable service of any station is that obtained from ground waves. The extent of the areas of ground-wave service is determined not only by the transmitter power, by the frequency and type of antenna, and by the ground conductivity in the area, but also by the interference to the desired signal caused by atmospheric noise, man-made noise, other stations on the same or adjacent channels, or, under certain conditions at night, by the fading and distortion caused by a mixture of ground wave and sky wave. Since the intensity of these limiting factors varies widely from moment to moment and from night to night, the area of satisfactory service also varies.

Since the strong steady ground-wave service is in general available only within a relatively short distance of the transmitter, a considerable part of the country lies outside such areas. This is particularly true at night, when strong sky-wave signals from distant stations on the same channel cause considerable interference in many instances and thus reduce the effective service area from its daytime value. At night in these areas use can be made of any interference-free sky-wave signals for service, but, because of its wide variation in intensity and occasional periods of signal distortion, such service is considered in general less desirable than ground-wave service. The extent of the area of satisfactory sky-wave service depends upon the interference to the desired signal caused by atmospheric and man-made noise, other stations on the same and adjacent channels, and under certain conditions by the fading and distortion caused by a mixture of the station's own ground- and sky-wave signals. In view of the wide variation in sky-wave-signal level from night to night, the area in which satisfactory listening can be had on any one night will vary greatly.

The zone in which fading is first encountered is at that distance where the sky wave becomes of such intensity as to interfere with the ground wave. The primary range of the station may be extended either by increasing the strength of the ground wave or by decreasing the strength of the sky wave (see Section 10, article 23). If the receiver is located within the incipient fading distance an increase in transmitting power above the noise level improves reception and increases the primary range. In the fading zone, however, an increase in power beyond that required to produce an average field intensity sufficient to override noise produces no further increase in service area. This is because the strength of the ground wave and that of the sky wave are being increased simultaneously and thus the relationship between them is maintained constant. It is therefore evident that with the low-powered transmitter the primary range is limited by the field intensity falling below the prevailing noise level at the receiving point. With the high-power transmitter the primary range is more likely to be limited by fading and attendant objectionable phenomena since the ground wave will usually be strong enough to override noise out to and beyond the point where fading begins. For detailed calculations of broadcast station range see Section 10, article 24.

F-M BROADCASTING COVERAGE. Although some service may be provided by tropospheric waves, the service area of a f-m broadcasting station is considered to be only that served by the ground wave. The extent of the service area is determined by the point at which the ground wave is no longer of sufficient intensity to provide satisfactory reception. The field intensity necessary for service in city, business, or factory areas is generally considered to be 1000 microvolts per meter. In rural areas, on the other hand, 50 microvolts per meter is generally believed to be sufficient for good reception. These figures are based upon the usual noise levels encountered and upon the absence of interference from other stations.

The ground-wave-signal range of a f-m broadcasting station is a function of the heights of the transmitting and receiving antennas, the gain of the antennas, the transmitter power, the frequency, the ground conductivity, and the dielectric constant. The service area of a f-m station, just like that of standard a-m broadcasting stations, may be accurately calculated by known methods (see Section 10, article 20). A detailed study of the service areas possible with f-m broadcasting stations develops these facts:

1. The service area of approximately half of all United States low-power (100- and 250-watt) standard a-m broadcast stations could be increased by going to frequency modulation.
2. Standard broadcast stations in areas of poor soil conductivity would benefit by a change to frequency modulation.
3. Standard broadcast stations having frequency assignments in the high-frequency end of the band would gain by a change to frequency modulation.

INTERFERENCE TO BROADCAST SERVICE. The strength of the electric field produced at the receiving location depends on many factors such as the power of the transmitting station, the nature and efficiency of the antenna system, the distance involved, the nature of the intervening terrain, and in some cases the time of the day and the season of the year. At a particular receiving location, in addition to the electric field strength produced by the desired broadcast station, other electric fields will exist which may hamper or prevent reception. It is not the absolute electric field strength produced by the desired broadcast station which determines whether reception will be satisfactory; it is the ratio of the desired field strength to the predominating interfering fields, coupled with the ability of the receiving set to discriminate against those interfering fields, which determines the success of the reception.

Interfering fields may arise from atmospheric disturbances (static), from industrial electrical interference, and from stations operating on the same or different channels.

The intensity of the atmospheric noise is not constant throughout the radio-frequency spectrum. At night it varies inversely with the frequency; daytime atmospheric noise varies approximately inversely as the square of the frequency. The magnitude of the noise depends upon the geographical location of the receiving point, the season of the year, and the conditions existing at the receiver.

Industrial electrical interference produced by the operation of non-radio electrical devices is, on the average, inversely proportional to the radio frequency. There are also present within the receiving apparatus, itself, sources of noise that require consideration: resistor noise, tube noise, contact noise, and noise associated with the tube power supply. In general, this receiving-set noise is independent of the radio frequency to which the receiver is tuned.

CONTINUITY OF SERVICE. One of the prime prerequisites for the successful operation of a broadcasting station is the absolute continuity of service throughout the broadcast day. The necessity for such operation arises from the keen competition among the many stations in this country and the psychological reaction of the average listener to an interruption in the program service.

This requirement imposes a severe responsibility upon the equipment and on the maintenance crew of a broadcasting station, inasmuch as the stations at the broadcasting centers and those associated with the major chains operate from 16 to 18 hours continuously every day of the year. In planning a broadcasting station many precautionary measures must be taken and suitable devices must be provided to permit the instant isolation and replacement of any equipment that becomes defective during the course of operation.

18. FIDELITY REQUIREMENTS OF BROADCAST SYSTEM

A high-quality broadcasting system is one which acoustically transports the listener in fancy from his loudspeaker to the studio or auditorium. It must be free from frequency, non-linear, and velocity distortion. It must not introduce extraneous sounds of annoying nature or distracting magnitudes.

TONAL RANGE. Complete freedom from frequency distortion implies that the system should be uniformly responsive over the entire range of audible frequencies. The audible range of the ear depends upon a great many factors (see Section 12, article 2), the extreme limits being in the neighborhood of 16 to 16,000 cycles per second—some 10 octaves. In practice, few persons can hear this extreme range at the listening levels normally used in the home (65–75 db above the acoustical reference level of 10^{-16} watt per cm^2). Furthermore, studies seem to indicate that few care to hear this extreme range when listening to broadcast music.

Several kinds of investigations have been undertaken to obtain data that would be of assistance in determining the optimum tonal range of a practical system. One series of tests was based on the acuity of hearing of the listeners. These experiments were concerned only with the physical ability to hear differences in band width; they disregarded the question of the enjoyment or esthetic appreciation of wider bands. The types of program material included a dance orchestra, two symphony orchestras, male speech, and a dramatic sketch. The observers were engineers who had had extensive experience in tests of program quality and were considerably more critical than the average radio listener. It was found that a change of band width from 15 to 8 kc had to be made to be as readily detected as a change from 8 to 5 kc. These changes, for speech, are just sufficient to have an equal chance of being detected by listeners having experience in such tests.

Changes in band width were found to be about twice as readily detected with music as with speech. Thus, for music, the changes that were just discernible half of the time were found to be 15 to 11 kc, 11 to 8 kc, 8 to 6.5 kc, and 6.5 to 5 kc.

In another kind of test the tonal range preferences of a cross-section of radio broadcast listeners were studied. As contrasted to the studies that have been made to determine the ability to distinguish between different band widths, this undertaking ascertained the tonal range that the average listener considered most pleasant, that is, the method of reproduction the listener would select in his home when listening for enjoyment. Classic and popular music, male and female speech, piano, and mixed voices with sound effects were employed. Every possible precaution was taken during the tests to remove any possibility of factors other than tonal range from influencing the listeners. A noise-free, essentially distortionless system was used, and the reproduction level was adjusted in accordance with the listener's desires.

In these tests the cut-off of both the low and the high frequencies was gradual, in keeping with the type found in actual radio receivers. It was found that listeners preferred a tonal range whose upper frequency limit was down about 3 db at 5000 cycles, about 20 db at 8000 cycles, and about 30 db at 10,000 cycles with respect to the mid-range frequencies. (The experiments did not test the preference for different rates of cut-off but, rather, for different tonal ranges all having cut-offs of about the same rate as the one mentioned.) Most listeners preferred a limited tonal range to a wider one even when told that one condition was representive of "low fidelity" and the other of "high fidelity."

In practice, broadcast transmitting systems are designed to provide uniform transmission over a wide range of frequencies. A-m broadcast transmitters, for example, are capable of covering the audio spectrum from 50 to at least 10,000 cycles, with negligible variations. F-m broadcast transmitters cover a still wider band, extending to at least 15,000 cps. On the other hand, except for a few isolated instances, commercially available receiving sets are not capable of faithfully reproducing anywhere near this range of frequencies.

Intercity network wire facilities having a very uniform frequency characteristic, particularly at the higher frequencies, can be secured, but their general use is a matter of economic consideration. For all practical purposes, the overall frequency-response characteristics of a complete broadcasting system is limited by the wire line characteristics.

The rate at which high-fidelity receiving equipment is put into service will, to a great extent, influence the employment of better wire line facilities between studios and radio stations.

DYNAMIC VOLUME RANGE. The dynamic volume range of a sound source of varying intensity is the ratio of the loudest sound produced to the minimum sound that is distinguishable. In broadcasting and sound recording, the loudest sound intensities are usually experienced with symphonic orchestras or special sounds such as explosions, gunfire, and factory noises (see Section 12). The minimum audible sound intensity is a function of the residual noise level.

As noted above (article 1), the maximum sound intensities encountered in studios is about 95 db for music. Furthermore, it was stated that room noise levels of 25 db are generally considered satisfactory. Thus it is evident that the maximum dynamic range likely to be encountered in original performances is about 70 db (excluding special sounds which may reach any intensity). This is a somewhat wider range (about 10 db) than can be accommodated by most complete sound-reproducing systems. It is considerably wider than the range that listeners prefer.

Very few listening environments are capable of making full use of even a 60-db dynamic range. In the home, for example, the average listening level is between 65 and 70 db above the acoustical reference. The residual noise level, on the other hand, is 43 db in the average residence, and in only 1 per cent of the homes is it as low as 30 db. Thus, even in the quietest suburban homes the noise level is about 40 db below the average listening level and in the average home about 30 db down. However, the source of the noise is not likely to be in the same direction from the listener as the radio receiver. Consequently, the benefits of binaural hearing (article 1, above) will assist the listener in partially disregarding room noise. Nevertheless, at the average listening level, a 60-db dynamic range cannot be fully exploited by the listener even in the quietest homes.

As a corollary to the question of dynamic range, it has been found, by studies in which a cross-section of broadcast listeners participated, that listeners prefer to hear music and speech at about the same *peak* levels (as read by a standard volume indicator, see article 4, above). It was also found that the limit of the range of *peak* volume levels tolerated by the largest number of listeners is approximately 8 db (4 db above and below the average volume level of the program). Even within this 8-db range it appears that changes in volume level are less annoying when made gradually. The 8-db limit refers to the range of *peak* or maximum volume levels, *not* to the range of minimum and maximum sound intensities or "dynamic range." (It is important that this preferred range in *peak* levels not to be confused with dynamic range, which was discussed in the opening paragraphs of this section.)

It was also found that, regardless of the absolute sound intensity at which the listener operates his radio, he still prefers an even peak intensity level. This is true whether he is listening to variety, drama, narrative, or musical programs. The peak intensities of the main program material (but not necessarily background effects) must not fall more than 8 db below the maximum peak level; otherwise the conditions for easy listening are violated.

BIBLIOGRAPHY

Chinn, H. A., and P. Eisenberg, Tonal-range and Sound-intensity Preferences of Broadcast Listeners, *Proc. I.R.E.*, September 1945, p. 571.
Chinn, H. A., and P. Eisenberg, New Broadcast Program Transmission Standards, *Proc. I.R.E.*, 1947.
Chinn, H. A., and P. Eisenberg, Influence of Reproducing System on Tonal-range Preferences, *Proc. I.R.E.*, May 1948, p. 572.
Gannett, D. K., and I. Kerney, Discernibility of Changes in Program Band Width, *Bell Sys. Tech. J.*, January 1944, p. 1.
Seacord, D. F., Room Noise at Telephone Locations, *Elec. Eng.*, June 1939, p. 225; June 1940, p. 232.

POLICE RADIO

By H. F. Mickel

Police radio systems may be divided into two major classifications from an operational standpoint. The simplest type of system, designated "one way," permits communication in one direction only, from the headquarters station to mobile units. This type of operation requires a land station transmitter for the headquarters location and a receiver for each mobile unit. The "two-way" system provides communication from the headquarters station to mobile units and from mobile units to headquarters. The equipment required for a two-way system consists essentially of a land station transmitter, one or more land station receivers, and a mobile receiver and transmitter for each two-way vehicle. In certain installations where all equipment is on the same frequency, or where mobile transmitters are equipped for two-frequency operation, a "three-way" system is evolved permitting car-to-car communication in addition to the two-way previously described. Essentially all systems which have been placed in service since 1942 are of the two-way or three-way type.

Police radio systems may also be divided into two major groups on the basis of the kind of equipment used. Prior to 1940, practically all police radio installations employed a-m apparatus. Since that date, the vast majority of systems have made use of f-m equipment. The only activity in the installation of a-m apparatus is confined to the replacement or expansion of existing systems. New systems, almost without exception, are of the f-m variety.

The scope and complexity of police radio systems vary with the requirements of each particular installation. A small municipality may operate a single headquarters station and a small number of mobile units. If conditions at the system control point are not desirable for the local installation and control of the land station, a location providing advantages of increased elevation and improved noise-level conditions may be selected for this equipment. This requires remote-control apparatus at the control point and the interconnection of the control and station locations by means of wire line.

In larger cities where a great number of mobile units and a considerable coverage area are involved, fixed station equipments may be located at several points with individual control from one central dispatching station or from separate precinct control points.

In many cases, individual receivers of the type used in land stations are located at several advantageous points throughout a city with their outputs feeding back to the control station or stations over wire line. This greatly increases the talk-back range of the mobile unit to the fixed station.

State police systems normally involve a multiplicity of land station transmitters and receivers located strategically to cover desired troop or patrol areas.

A number of state and large city systems also incorporate CW telegraph stations for zone and interzone point-to-point communication.

In some instances, radio relay equipment is used for the control of remotely located stations, particularly where topographical conditions necessitate such remote installation of land station equipment and render impractical the use of wire line interconnection to the control station.

19. FREQUENCIES

The first police radio systems operated on frequencies just above the standard broadcast band with channels in two portions of the spectrum (1610 kc–1730 kc and 2326 kc–2490 kc) assigned for this purpose. The general plan was to place state systems in the lower band and city stations on the higher channels. A number of these systems are still in operation, and essentially all apparatus employing these frequencies is for land-station-to-mobile communication. These, doubtless, will be gradually replaced by equipment operating on higher frequencies.

The next portion of the spectrum assigned for police operation was the 30.1–40.1 megacycle band (now 30–50 megacycles). It was found that these frequencies possessed many operational advantages for police work: reduction of interference between stations, reduced atmospheric and man-made interference, lower transmitter power output requirements, and, probably most important of all, the ability to produce practical equipment for mobile talk-back operation. It was in this band that f-m apparatus for mobile communications made its appearance with its many attendant advantages.

During World War II the use of still higher frequencies by the Armed Forces disclosed many characteristics which pointed toward the desirability of their use for police communications systems. Some background of experience gained in the use of 116- to 118-megacycle equipment for relay purposes in police systems, starting about 1940, also gave added weight to this belief. Experimental work encompassing frequencies in the 160-megacycle area revealed very favorable performance characteristics for police use. Accordingly, the Federal Communications Commission has set aside channels in the 152- to 162-megacycle band for the police services. From actual data covering a representative number of systems in normal operation, it appears that sky-wave or "skip" interference, often causing considerable trouble in the 30- to 50-megacycle band, is greatly reduced in the 152- to 162-megacycle frequencies. Smaller antennas are a natural and desirable result of the use of 152- to 162-megacycle channels with the further advantage that high gain and directional antennas for land stations are entirely feasible at these frequencies. Atmospheric and man-made interference is still further reduced as compared to the 30- to 50-megacycle band. Coverage within the normal service area of a 152- to 162-megacycle system is more complete, with fewer dead spots, than with lower frequencies. It appears, also, that still lower transmitter powers will give satisfactory results.

The Federal Communications Commission has also made certain shared assignments for police service in the 72- to 76-megacycle band. Tests conducted on the 30- to 50-megacycle, 72- to 76-megacycle, and the 152- to 162-megacycle bands indicate that the extreme coverage range decreases somewhat as the frequency rises but that improved blanket coverage within the useful range is achieved as the frequency is increased. The 30- to 50-megacycle channels are characterized by a greater degree of "bending" of the transmitted signal and, for that reason, seem better suited for applications where greater coverage distance is required, particularly in hilly or mountainous terrain.

It appears, therefore, that 30- to 50-megacycle channels are best suited for state and county police systems and 152- to 162-megacycle frequencies for municipal use. The Federal Communications Commission has also provided bands for police service in various portions of the spectrum from 450 megacycles to 30,000 megacycles. Complete information regarding all allocations of frequencies for police use may be obtained from the May 6, 1949, issue of the Federal Register or from the new FCC Rules and Regulations, when published.

20. POWER AND RANGE

There is no fixed formula for the absolute determination of coverage which may be expected in a police radio system. Local conditions of terrain, antenna elevation, and noise level are some of the variables that influence such coverage. Transmitter power output is also a factor, particularly in the 1610- to 1730-kc and 2326- to 2490-kc bands. However, in the higher-frequency channels, antenna elevation and noise-level conditions are more influential than transmitter power. Table 1 indicates the normal limit of transmitter power in the various frequency bands (plate power input to final stage).

Since communications range is a function of so many variable factors, actual experience in the planning and installation of police radio systems is the most reliable means of predicting results. A single land station installation will afford satisfactory two-way communication for an average city county, provided that the antenna site is carefully selected from the standpoint of elevation and noise level. Two-way range up to 50 miles with 30- to 50-megacycle equipment is not uncommon with modern apparatus properly installed. A slight decrease in range may be expected with 72- to 76-megacycle and 152- to 162-megacycle equipment.

Table 1

Frequency, megacycles	Land Station Power, watts
1.61–3.0	2000
25–100	500
100–220	600
Above 220	To be specified in authorization

21. EQUIPMENT

Land station transmitting equipment is of conventional design using standard circuits and tubes. F-m equipment is normally of the phase-shift type. The same applies to mobile transmitters.

Land station and mobile receivers are of the superheterodyne variety and are fixed tuned to the assigned operating frequency. Squelch circuits are provided to quiet the receivers when the associated carrier is not on the air.

Crystal control has become standard for all transmitting and receiving equipment designed for police service.

Power for land station equipment is normally obtained from the regular public utility service. Frequently gas-engine generating equipment is provided for emergency operation in the event of failure of the regular power source.

Mobile equipment uses the car battery as the primary power source with either vibrator or dynamotor units for high-voltage d-c supply.

Since vertical polarization has become standard in police service (30 megacycles and up), mobile antennas are of the vertical whip type for either side or roof top mounting. Land station antennas vary somewhat in design but are usually of the J, coaxial, or ground plane type.

SECTION 17

TELEPHONY

BY

JOHN D. TAYLOR

TELEPHONY

By John D. Taylor

TELEPHONY is the art of electrically transmitting speech between two oi more points. Telephone facilities are also used for many other purposes, such as the transmission of broadcasting and public-address programs.

Transmission of speech and other forms of intelligence is accomplished over wire circuits or through the air (radio) or by a combination of both mediums.

The telephone circuit fundamentally consists of a device (transmitter) for transforming speech sounds into electrical currents, which traverse a connecting medium (line or channel) and react in another device (receiver) in such manner as to convert the electrical currents into the original speech sounds.

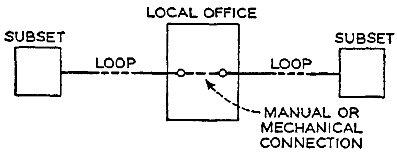


FIG. 1. Connection between Two Subscribers in the Same Office (Courtesy Bell System)

Switching arrangements of various types and capacities, either manual or mechanical,* are necessary to connect local or toll telephone circuits together, and a number of auxiliary circuits, in addition to the talking circuit, may be employed for a given connection, depending upon the types of systems involved and the length of the connection. The interconnection of two subscriber lines (loops) in the same office is usually quite simple, but for lines in widely separated offices the complete interconnecting circuit and associated apparatus may be very complex.

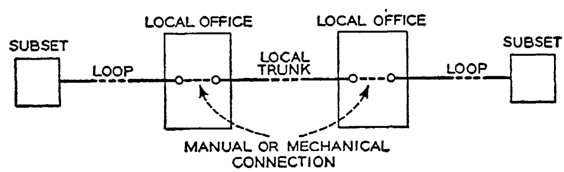


FIG. 2. Connection between Two Subscribers in Different Offices in the Same City (Courtesy Bell System)

Representative types of telephone connections between two subscribers are shown schematically in Figs. 1, 2, 3, and 4.

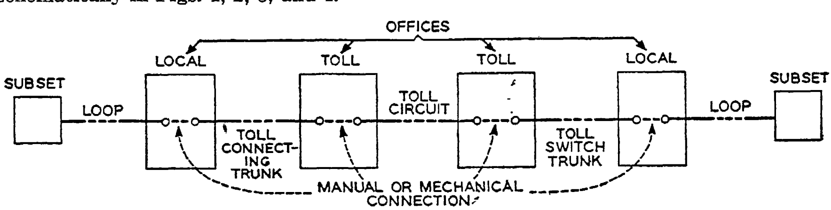


FIG. 3. Connection between Two Subscribers in Different Cities (Courtesy Bell System)

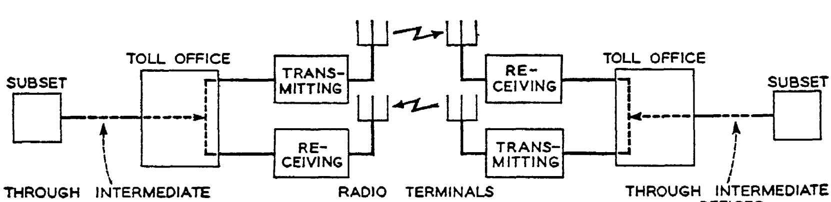


FIG. 4. Connection between Two Subscribers in Different Countries via Radio Channel (Courtesy Bell System)

* The word "mechanical" is used in this section in a broad sense to include all forms of non-manual switching.

In telephone practice the various facilities naturally fall under four main headings: (1) central-office equipment; (2) land and buildings; (3) telephone lines; (4) substation equipment.

CENTRAL-OFFICE EQUIPMENT

Central-office equipment, in general, embraces the various switching arrangements, including auxiliary units of equipment, which are necessary for the interconnection, disconnection, control, and supervision of telephone facilities. Usually, the larger the telephone system, the more intricate are the equipment requirements.

The evolution of telephone service has been from magneto to manual common-battery to mechanical common-battery operation. All these types of operation are now in use, but the trend in present-day engineering is to mechanize telephone service in order to obtain greater speed, ease, and efficiency of operation and to avoid higher operating costs.

Central-office equipment includes all types of switchboards, both manual and mechanical, switch frames and panels, terminating frames and racks, toll terminal equipment, testing units, power plants, and many auxiliary pieces of apparatus. The equipment is housed in suitable buildings, and each entire assembly is known as a central office.

Auxiliary circuits and apparatus, such as alarms and indicators, both visual and aural, designed to call attention to certain operating conditions, monitoring and supervisory circuits, timing and recording devices, emergency power and ringing circuits, test circuits, and many other devices, necessary for the proper operation of central offices, are common to all systems to a greater or less degree, depending on the type and size of system.

1. MANUAL SYSTEMS AND OPERATION

Manual systems include both magneto and common-battery systems, in which telephone operators manually establish and supervise connections at switchboards, using flexible cords or keying units.

Magneto operation, first employed in the United States, requires operation of a magneto or hand generator (associated with magneto telephone sets) by the subscriber to signal the operator for connections and disconnections, and the provision of local battery (dry cells) at each telephone. Present practice also provides for hookswitch signaling (with limitations) by the subscriber on magneto lines, if desired, similar to common-battery operation.

Common-battery operation provides for the subscriber to signal the operator by removing his handset * from or replacing it on the telephone set hook,* direct current for both signaling and talking being supplied to the telephone set from the central-office battery over the subscriber line.

Manual switchboards are of several types and are made by a number of different manufacturers for interconnecting toll, trunk, and subscriber line circuits.

Magneto switchboards employ simple cord and line circuits but, in general, provide the least desirable telephone service from the standpoint of speed and ease of operation. These boards are now built with capacities of up to 200 subscriber lines and are used principally in small offices, where the majority of the terminating lines extend into rural areas and have relatively high energy losses. Even here the present tendency is toward mechanical equipment in new installations and replacements.

Typical full magneto switchboard circuits of the latest type, for a board having a capacity of 150 lines and 15 cord circuits, are shown in Fig. 1.

FULL MAGNETO OPERATION of such a switchboard is as follows:

In placing a call, the subscriber turns the hand generator crank at his telephone, sending 20-cycle current over his line and operating the switchboard drop (or line lamp circuit, if furnished). The subscriber then removes his handset from the hook and listens for the operator to answer. The operator inserts the answering plug of an idle cord circuit in the line jack associated with the operated drop, opens her listening key, and requests the called number. Assuming that this number is also a local subscriber's line in this office, the operator inserts the calling plug (associated with the cord circuit being used) in the called-number line jack and operates her ringing key, using code ringing as may be required, to signal the called station. If ringing power is supplied to the switchboard by a hand generator, the generator crank must also be turned by the operator while operating the ringing key. The called station bell is actuated by the 20-cycle current sent out over

* The word "handset" is intended to include the older-type telephone receiver and the word "hook" to include the newer-type telephone-set cradle.

the line, and the operator awaits the called subscriber's answer with her listening key open; when she hears the subscriber answer she closes her key and disconnects her telephone set from the connection.

On completion of the call both subscribers place their handsets on the hook, the calling subscriber turns his generator crank, operating the answering cord ring-off drop and sig-

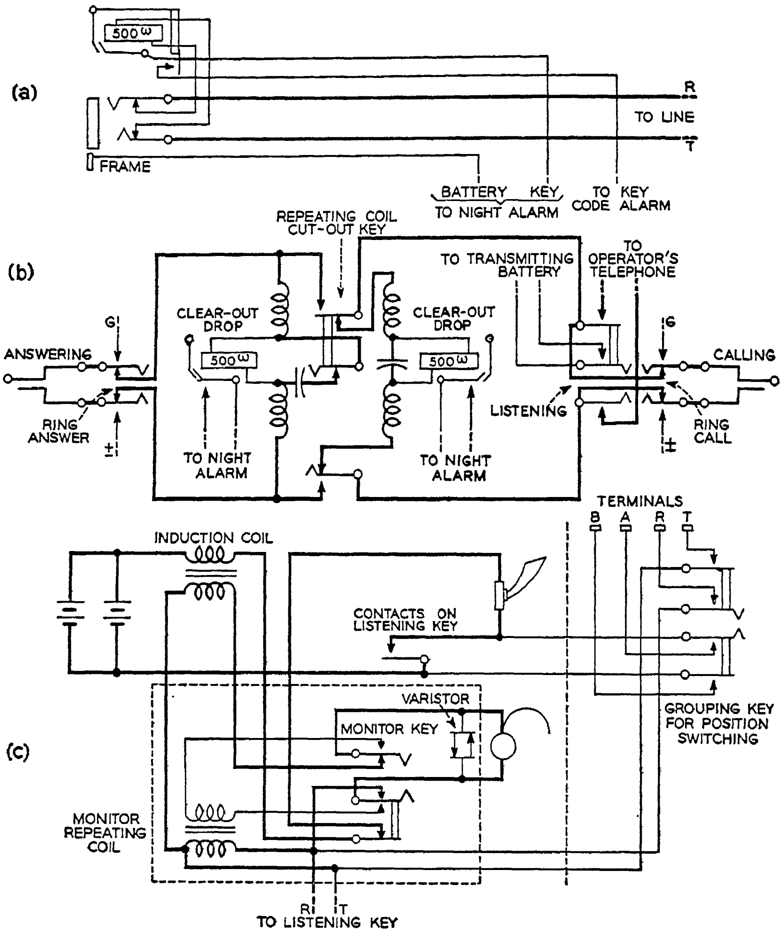


FIG. 1. Full Magneto Switchboard Circuits—Magneto Signaling by Subscriber (Courtesy Stromberg-Carlson Co.)

(a) Line circuit with drop arranged for code alarm.
(b) Cord circuit with repeating coil, double clear-out drops, and repeating coil cut-out key.
(c) Operator's circuit, including monitor, varistor, and grouping key circuits.

naling the operator that the conversation is completed. She then removes the plugs from the line jacks and restores the cord circuit to its idle position.

Central energy (common-battery) type telephones on magneto lines may be made to operate successfully, using a line circuit, as shown in Fig. 2. This circuit provides service to a subscriber in a magneto office similar to common-battery operation, but such service is limited generally to short town lines, not exceeding about 225 ohms conductor loop resistance. The subscriber removes his handset from the hook, sending a surge of direct current through the repeating coil windings (line side), the line, and the telephone set. This surge induces a surge of current in the drop windings of the repeating coil, which are in series with the line drop. The line drop is operated and the call is handled from that point in the same manner as other full magneto calls.

Magneto toll lines are terminated at magneto switchboards in the same type of line circuit, and the operating is similar to that for local circuits. In the local switchboard, two pairs of cord circuits are arranged so that the repeating coil can be removed by operating a key on toll connections, in order to reduce transmission loss, assuming that circuit

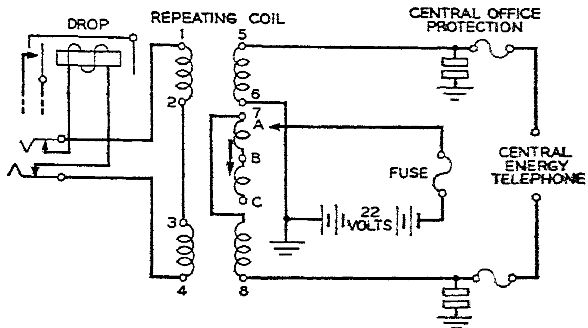


Fig. 2. Magneto Switchboard Subscriber Line Circuit Arranged for Common-battery Signaling and Talking by Subscriber (Courtesy Stromberg-Carlson Co.)

noise is satisfactory without the coil for a particular connection. Figure 3 shows typical loop and simplex dial trunk and cord circuits for use at magneto switchboards, which have trunks to a mechanical office.

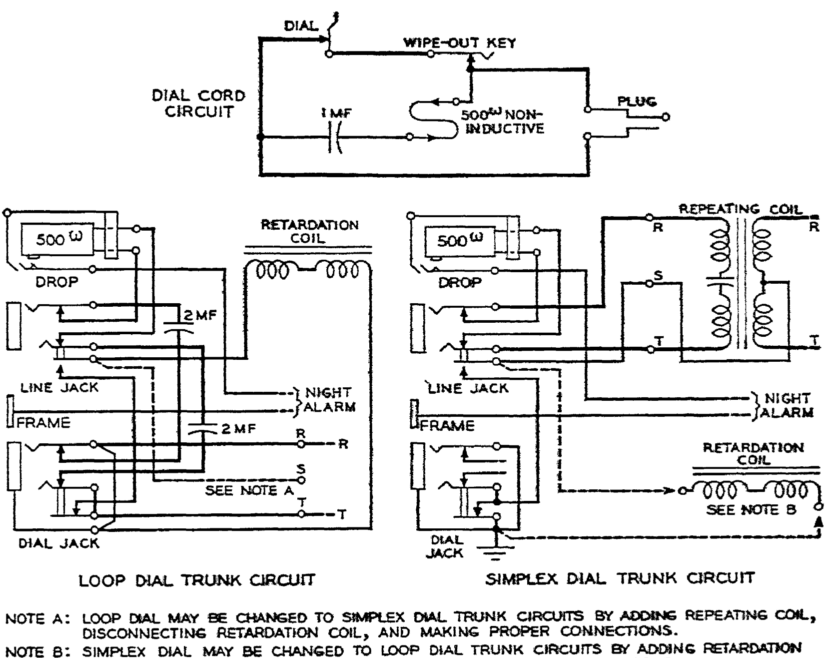


Fig. 3. Magneto Switchboard—Dial Trunk and Cord Circuits (Courtesy Stromberg-Carlson Co.)

COMMON-BATTERY SWITCHBOARDS are made in a variety of types and capacities, both single and multisection (multiple), to meet service requirements and are widely used throughout the United States and foreign countries. The subscriber signals the switchboard operator by operating the hookswitch (or cradleswitch) of his telephone. However, in order to provide this more convenient and faster service, the subscriber switchboard line, cord, and auxiliary circuits are more complex than for magneto equipment.

Single-section common-battery non-multiple switchboards are used principally in the smaller towns, where magneto service is not adequate. In this type of board each subscriber's line appears in the switchboard jack field only once, since one operator can reach

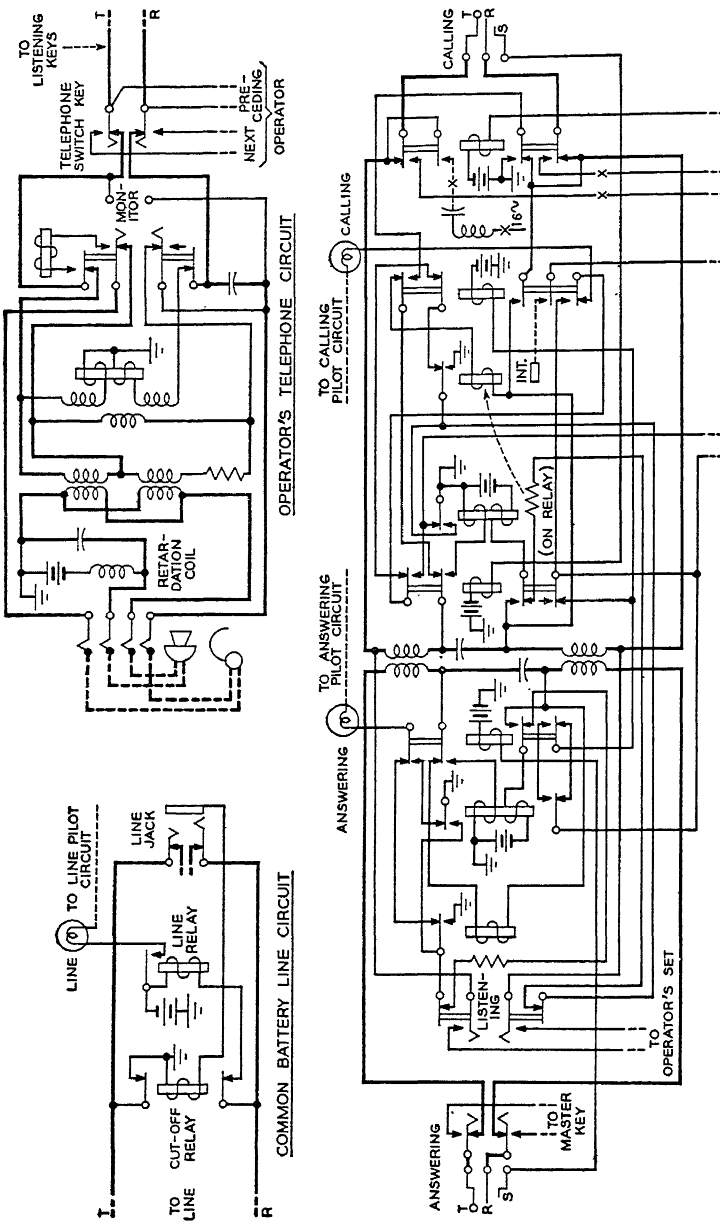


Fig. 4. Single Section Common-battery Switchboard Circuits—Common-battery Talking and Signaling (Courtesy Kellogg Switchboard and Supply Co.)

any jack in the board. Figure 4 shows schematic circuits of a typical board of this type. The capacity of such a board is up to 200 subscriber lines, 30 toll or rural lines, and 16 universal cord circuits. The capacity may be doubled by operating two such boards adjacent to each other.

Multiple common-battery switchboards are designed for single-office and multioffice cities, where single-section boards are inadequate to meet service requirements. The capacities of this type of board range from about 600 to about 10,000 lines, thus limiting the capacity of a single office to about 10,000 lines. For the large cities, requiring more than 10,000 lines, more than one office, each with its own switchboard, is necessary.

Single-office multiple common-battery switchboards are assembled by sections in one or more line-ups, each section being identically equipped with jack fields and cord circuits. The number of sections in an office varies from two to twenty or more, depending upon

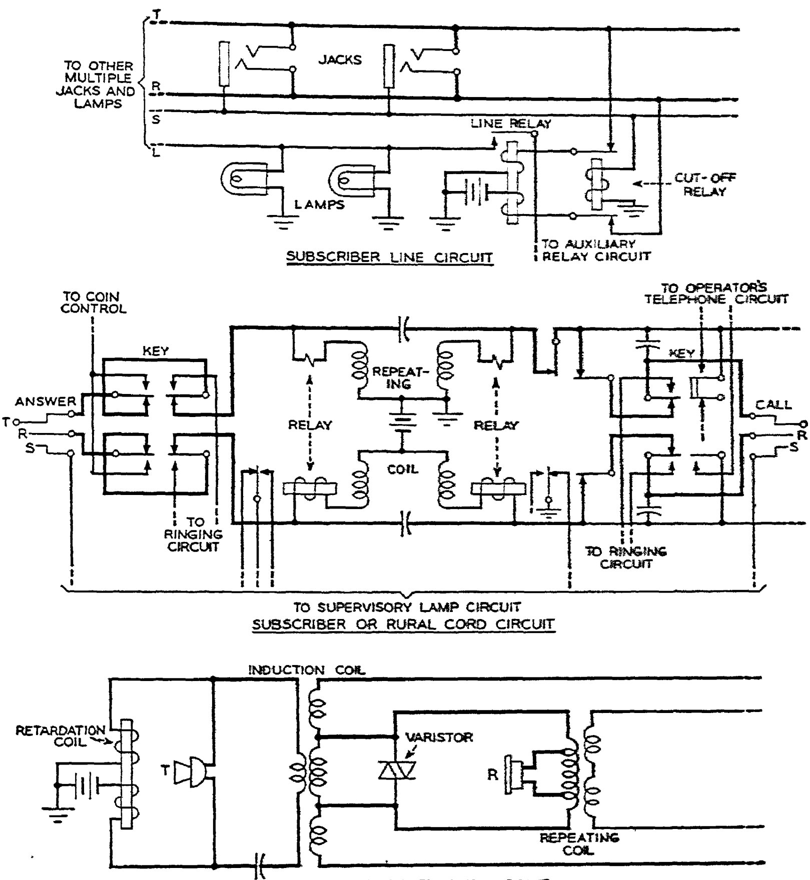


FIG. 5. Multisection Common-battery Switchboard Circuits—4000-line Common-battery Talking and Signaling (Courtesy Bell System)

the number of subscriber lines served and the traffic load. Each section of switchboard provides for from one to two and two-thirds operators and from three to eight panels, in which the multiple jack and lamp strips are mounted. Each subscriber line has one multiple jack with associated line lamp in each section, although in some of the older boards only one answering jack was provided per line. Thus, jack 100 and its associated lamp in the first section in the line-up are cabled to jack 100 and its lamp in the second section, and similarly throughout the board.

Since each subscriber line terminates in the jack and lamp circuit corresponding to his number, when the subscriber signals the operator to place a call all the line lamps associated with his line throughout the board light and may be answered by any available operator, but by only one at a time. When one operator answers a call, all jack sleeves associated with that particular line have potential placed on them, which causes a click in the ear of any operator who touches the tip of her plug to the sleeve of any jack associated with that line. This click warns the operator that the line is busy. The number of line lamps which are permitted to light on any one line is usually limited to five but may be less, depending on traffic loads and calling rates. Figure 5 shows the principal schematic circuits for a typical multiple common-battery switchboard with a capacity of 4000 subscriber lines, 360 toll lines or 720 outgoing trunks, and 17 cord circuits per position.

The provision of a trunking board and special arrangements of subscriber multiple in the various line-ups makes it possible to increase the capacity of the board to accommodate up to 5600, 7200, or 10,400 subscriber lines and to provide for a substantial complement of toll lines and trunks. However, when new central-office installations or sizable additions to existing manual boards are being considered, present practices require a careful study to determine the practicability and economies of employing mechanical operation, because of its many advantages, including integration with the general trend toward universal mechanized telephone service.

This board is capable of operating as a combined local and toll, local and trunk, or local, toll, and trunk board.

Multioffice multiple common-battery switchboards of modern design are similar to the single-office multiple board described above. Some of the older-type subscriber switchboards did not provide for multipling the line lamp as well as the multiple jack, so that the subscriber's lamp signal had only one appearance in the entire switchboard and answering time was considerably slower than with the multiple-line lamp arrangement.

INTEROFFICE TRUNKS are necessary in multioffice exchange areas to provide for extending a call from one office to another. In manual operation the calling-subscriber signal appears at the calling-subscriber switchboard (designated in trunking as the A board), the operator ascertains the called number and, either by the call circuit or straightforward trunking method, passes the called number to a terminating trunk board (designated in trunking as the B board) at the called office. An intermediate office (tandem) may be involved in establishing the trunk connection. The A operator connects the calling line to the selected outgoing trunk at the A board, the B operator at the called office connects the B end of the interconnecting AB trunk to the called B board multiple jack, and the ringing of the called subscriber automatically starts. The cycle of ringing usually consists of a 2-sec ringing interval followed by a 4-sec silent interval with d-c potential only impressed on the line. This cycle is repeated until the subscriber answers or the connection is taken down. When the subscriber answers in either the ringing or silent interval, relays in the B trunk circuit operate, disconnecting the ringing power and connecting the trunk circuit talking path through to the called subscriber. Upon completion of the conversation, a lamp disconnect signal appears before both A and B operators, in response to both the calling and called subscribers hanging up their handsets, and the connection is taken down.

Call circuit trunking is a procedure by which the A or toll operator passes a call to a B (or tandem) operator over a call (order wire) circuit, which is entirely separate from the trunk circuit being used for the call. When the A operator presses her call circuit key associated with the call circuit to the desired B (or tandem) office, she is connected directly to the distant operator's telephone set. After she passes the call to the distant operator, the distant operator assigns an idle trunk in the group between the two offices and the A and distant operators connect the trunk to the calling and called lines at their respective boards.

Straightforward trunking is now the generally accepted method in manual operation rather than the call circuit method, from which it differs in that the A operator selects the idle trunk to the called office. She then connects the calling subscriber to the trunk with an A cord circuit, causing a lamp to light at the distant operator's position. The distant operator connects her telephone set to the trunk by pressing a key, or the set is automatically connected and the A operator is so informed by hearing a two-tone signal on the trunk. The call is passed and the connection is established, and during the conversation the supervisory lamps at both A and B boards remain dark. When the subscribers hang up their handsets these lamps light and the operators disconnect.

2. MECHANICAL SYSTEMS AND OPERATION

Of the several types of mechanical switching systems now in operation, probably the Strowger (step-by-step) system, manufactured by the Automatic Electric Co. (and others under Automatic Co. patents) is most widely known. It was the first type employed commercially (in the year 1892) and is still used extensively today. Other well-known mechanical systems have been developed to meet the needs of the rapidly growing telephone industry, particularly the Relaydial (Stromberg-Carlson), Relaymatic (Kellogg Switchboard and Supply Co.), All-Relay (North Electric Manufacturing Co.), Panel Dial, and Crossbar

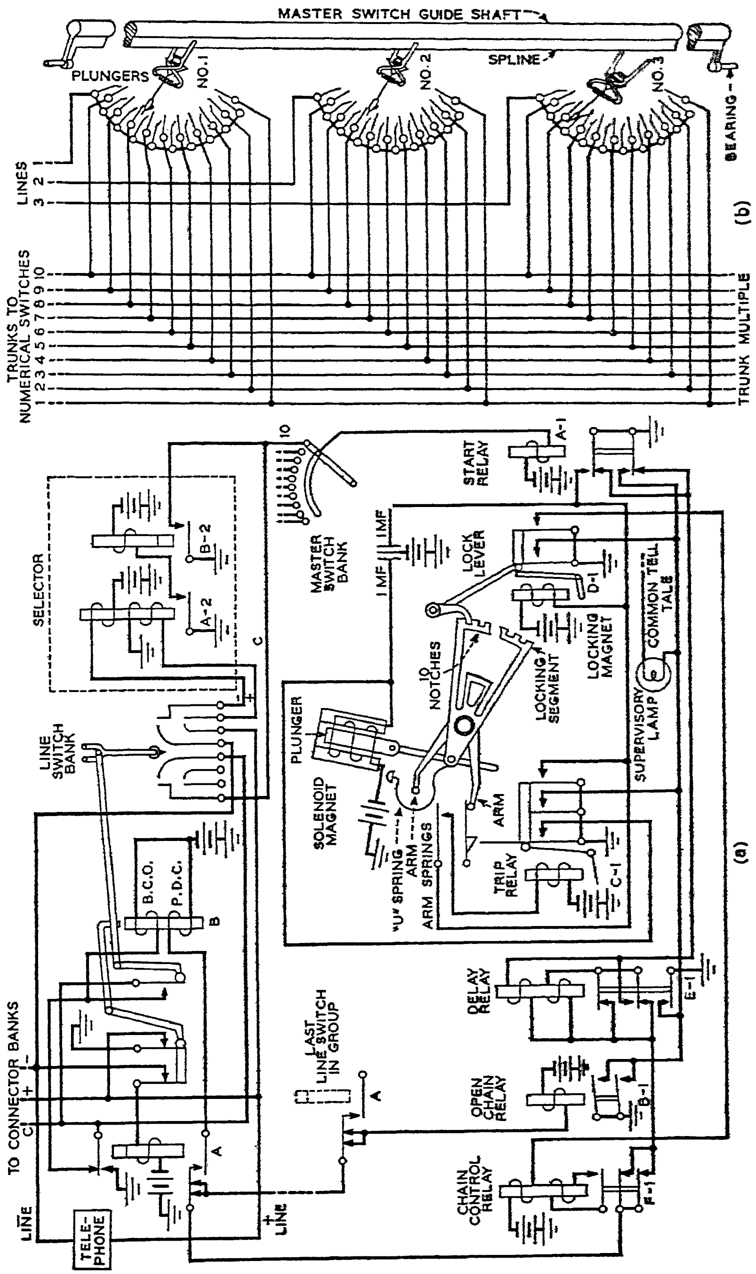


FIG. 6. Schematic Diagrams Showing (a) Self-aligning Line and Master Switch Circuit and (b) Relation between Plunger-type Line Switches and Trunks (Courtesy Automatic Electric Co.)

Systems (Western Electric Co.). All these systems have only one purpose, namely, to switch traffic quickly, accurately, economically, and in a manner satisfactory to the public, whether it be in a relatively small office or in the largest multioffice exchange area.

THE STROWGER SYSTEM employs the well-known step-by-step method of operation, so called because calls are advanced from the calling to the called subscriber step by step, as each digit of the called number is dialed by the calling subscriber.

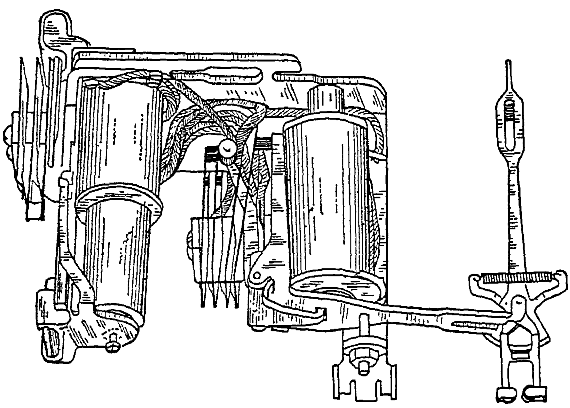


FIG. 7. Self-aligning Plunger Line Switch (Courtesy Automatic Electric Co.)

In step-by-step operation, the principal switching units involved in a connection between two local subscribers are a line switch or line finder, one or several ranks of selectors, and a connector. In addition, if the connection includes a trunk between two offices in the same exchange area, the outgoing end of the trunk will terminate in an impulse repeater.

The equipment which appears between the subscriber's line terminals and the first rank of selectors is classed as *non-numerical*, since it automatically functions as soon as the subscriber's handset is removed from its support and before any digits are dialed. Non-numerical switches are of two major classes—line switches and line finders. The line switch is individual to a telephone line and serves to extend the calling line to an idle selector or connector (forward selection), while a line finder is connected permanently to a selector or connector and serves to find the calling line (backward selection). Line switches are now seldom employed for public exchanges but are generally standard for private automatic exchanges of 100 lines or less.

The line switch may be of two types, *plunger* (10-trunk capacity) and *rotary* (10- or 25-trunk capacity).

The plunger line switch is a simple mechanism which automatically connects its calling line to any one of a number of trunks leading to *numerical* switches (the selector or connector, which are operated by dial pulses). Figure 6 shows a schematic diagram of the self-aligning plunger line switch and master switch circuit and of the wiring arrangement between trunks and line switches. Though only three line switches are shown in this latter arrangement, there may be from 25 to 100 such switches in one group. Figures 7 and 8 show views of the line and master switches.

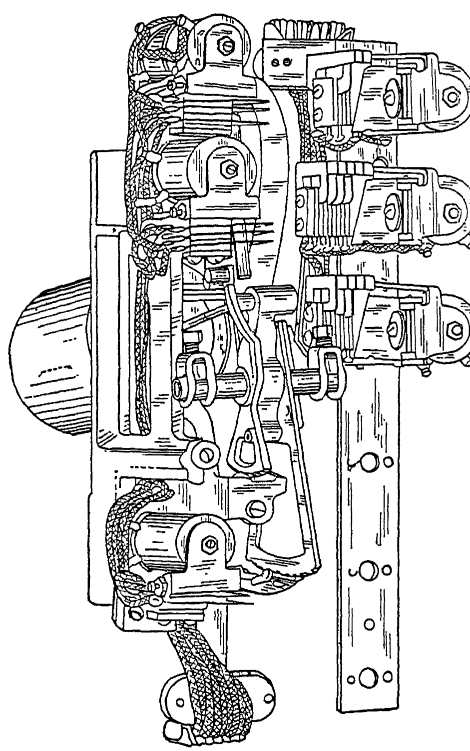


FIG. 8. Master Switch Associated with Self-aligning Plunger Line Switch (Courtesy Automatic Electric Co.)

When the handset is lifted at the telephone, the plunger is thrust into the line switch bank by operation of the A and B relays, closing the line and trunk spring contacts and extending the line through to the selector. The operation of the selector relays maintains the B relay and its plunger operated until the connection is released, the relays then returning to normal. Operation of the plunger starts operation of the master switch circuit, resulting in the moving of all idle line switch plungers opposite the next idle trunk. The self-aligning feature of the plunger reseats it on the master switch guide bar as soon as it is released. The connector bank terminal associated with the line is also made busy by the selector placing ground on the control lead.

The rotary line switch is a single-motion device, which may be associated with a telephone line for the purpose of extending the line to any one of a number of idle trunks. This switch has a shaft carrying wipers, which slide over bank contacts (arranged in a half circle) to which trunks to the numerical switches are connected. When the line and cut-off relays are energized by lifting the handset, the switch's driving magnet operates its

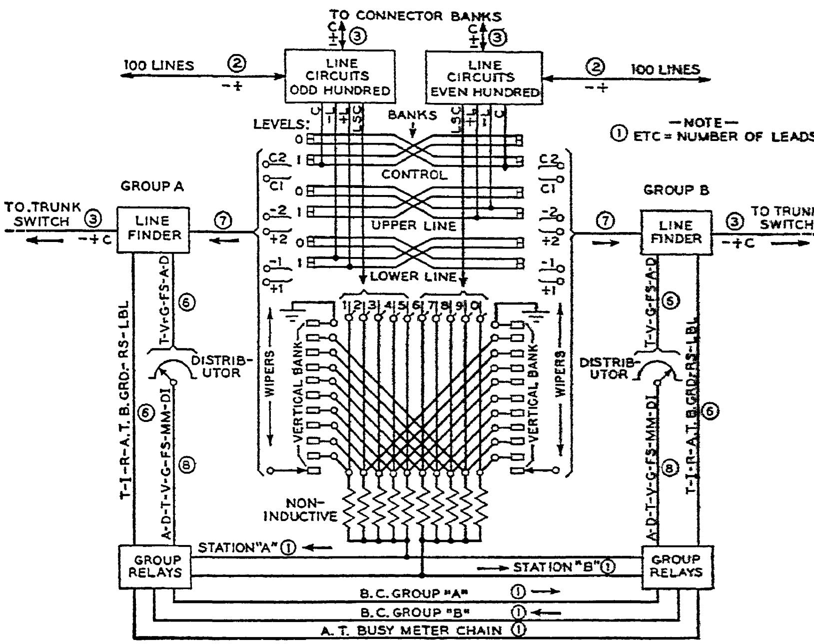


FIG. 9. Schematic Diagram of the 200 Line Finder Circuit (Courtesy Bell System)

armature against the action of a flat spring. At the end of the armature stroke the magnet circuit is opened, the spring forces the armature back, and the wipers are carried forward one step into the bank. This action continues until the wipers reach idle trunk contacts, where they remain until the next call is originated, when the wipers are moved to the next idle trunk. It is usually necessary to extend connections by means of wipers and bank contacts for several conductors which perform signal or control functions in addition to extension of the transmission path. Thus, each position of the wipers in the bank has from three to six or more bank contacts, which are simultaneously contacted by separate wipers when the selection is made. This type of switch is seldom used as a subscriber line switch since the plunger type is cheaper, but both operate satisfactorily.

Secondary line switches, rotary or plunger type, are now seldom used but were designed for larger step-by-step offices to combine the traffic from a number of primary groups of subscriber line switches and direct it to a relatively large common group of numerical switches capable of handling the combined traffic. Small trunk groups are less efficient than large trunk groups, and by employing secondary line switches the number of selectors required for the main trunk group need be only large enough to accommodate the peak load of the combined group instead of each small group requiring enough selectors to accommodate its own peak load, which usually will not occur at the same time as the peak loads of the other small groups.

The line finder switch seeks out the calling line from a group of subscriber lines connected to bank contacts and connects it to a trunk terminating in the first numerical switch of the switch train. The line finder switch permits the use of a simple, economical subscriber line circuit composed principally of a line and a cutoff relay comparable to those in common-battery manual systems. These relays function to *mark* the bank position of the calling line and to cause the allotted line finder to hunt the calling line.

As soon as the bank position of a calling line is marked by operation of the line relays, a relay of a common relay group (associated with each group of line finders) causes the proper line finder to hunt, first vertically and then horizontally, until the calling line is located, whereupon the finder connects its permanently associated first selector (or connector) to the calling line through the finder wipers and the bank contacts. The cutoff relay of the line circuit then operates to clear the line of unnecessary attachments and to free the common group of relays.

A number of line finders are grouped under the control of a single *distributor*, which preselects or allots the next line finder to be used in the next call as soon as the common relay group is released from the preceding call.

Line finder switches are of the two-motion (vertical and horizontal) type and usually have capacities for 50, 100, or 200 lines. In some small exchanges, line finders may be of the rotary (single-motion) type with capacities of 25 or 50 lines.

Figure 9 shows diagrammatically the switching arrangement between subscriber lines and trunks, and Fig. 10 a view of a line finder switch, for 200-line capacity. This switch has a group of relays mounted on a base, upon which is also mounted a frame supporting a shaft with ratchet mechanism for raising and rotating the shaft. The lower part of the shaft carries four sets of wipers (one single-conductor and three two-conductor), termed the *vertical, control* (upper bank), *upper line* (middle bank), and *lower line* (lower bank). The vertical and rotary stepping magnets and the release magnet (which permits the shaft to return to normal when the connection is released) are mounted within the switch frame.

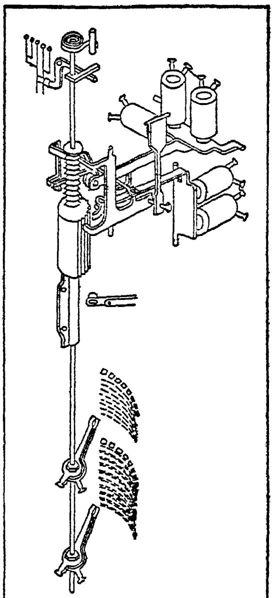


FIG. 10. Line Finder Switch for 200 Lines—Strowger System. (Courtesy Automatic Electric Co.)

A vertical interrupter (pulsing) circuit causes the vertical magnet (by its armature and pawl engaging the "vertical hub" ratchet) to elevate the shaft step by step to the marked level. A rotary interrupter circuit then causes the rotary magnet (by its armature and pawl engaging the "rotary hub" ratchet) to rotate the shaft step by step, until the marked control bank contact is engaged by the control wiper.

These motions cause the control, upper line, and lower line wipers to engage the corresponding contacts of their semicylindrical banks, each of which has 100 sets of contacts (10 levels and 10 sets per level). To the right of these banks is the vertical bank or commutator, comprising a single row of contacts, over which the vertical wiper moves until the marked level of the calling line is reached.

The release of the shaft is effected by the operation of the release magnet, which disengages the vertical, rotary, and stationary dogs from their ratchets, permitting the shaft to return to normal under spring and gravity action.

The impulse repeater is used in interoffice trunks in step-by-step exchange areas having more than one office. This repeater is required in the outgoing end of each trunk and functions (1) to make it unnecessary to provide a third (control) wire in each trunk, (2) to provide talking current to the calling subscriber, (3) to reverse battery to the calling subscriber when the called subscriber answers, and (4) to repeat dial pulses over the interoffice trunk so that the impulse circuit will not include both the subscriber's line and the interoffice trunk.

The two types of impulse repeaters are one-way and two-way. The first type is used at one end of one-way interoffice trunks; the second, at both ends of two-way interoffice trunks.

A diagram of the one-way impulse repeater circuit is shown in Fig. 11.

When the repeater is seized by the preceding switch, the A and B relays operate, and the B relay connects ground to the control lead C, to protect and hold the preceding switches in the train and avoid seizure by other switches. Relay B also operates relays A-1 and B-1 in the incoming selector at the distant office, establishing an impulse loop over the trunk. Relay A, responding to impulses from the calling subscriber dial, interrupts the impulse loop, according to the impulses received, thereby repeating these impulses over the loop.

When the called subscriber answers, operation of the back-bridge relay of the connector at the called office reverses the polarity of the current through the holding bridge (relay F) of the repeater at the calling office, causing relay F to operate. The operation of relay F causes relay D of the repeater to operate, which reverses the polarity of the current flow to the calling telephone, for the purpose of operating coin-collectors or message registers or of providing supervision of manual calls.

When the handset at the calling telephone is placed on its support, the relays release and the train of switches is restored to normal.

The selector is a switching device which became necessary for offices of over 100 lines. In a 1000-terminal system, only a single rank of selectors (first) is required, the first digit dialed operating the selector switch to connect to the desired hundred group of connectors.

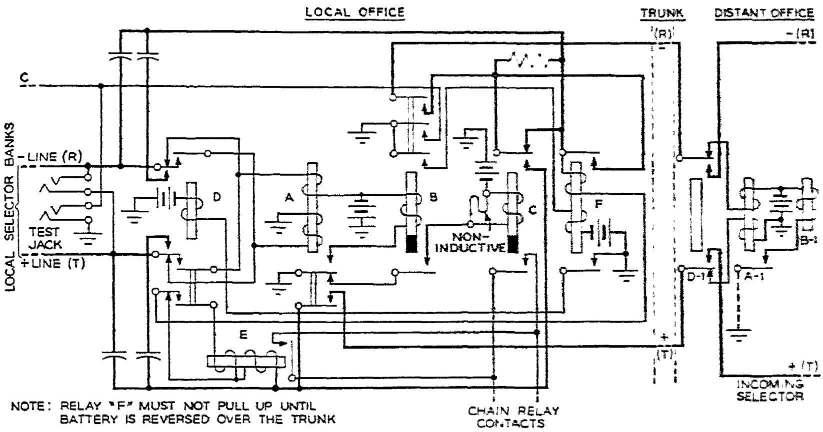


Fig. 11. One-way Impulse Repeater Circuit—Strowger System (Courtesy Automatic Electric Co.)

In a 10,000-terminal system, two ranks of selectors (first and second) are required. The first digit dialed operates the first selector to select the thousand group of trunks which terminate in second selectors, and the second digit dialed operates the second selector to select the desired hundred group of connectors. Thus, the selector is a numerical type of group-selecting, trunk-hunting two-motion switch, which requires but one digit for its operation.

Figure 12 shows a schematic diagram of the selector circuit. The selector has a group of control relays mounted on a base, which also supports a shaft and ratchet mechanism assembly for raising and rotating the shaft. The lower part of the shaft carries two sets of wipers, *control* (upper) and *line* (lower). The vertical and rotary (stepping) magnets and the release magnet are mounted within the switch frame. The bank contacts are in two groups of 100 sets of contacts each (10 levels and 10 sets per level).

When the selector is seized, it functions to hold all preceding switches in the train operated and guarded until the holding circuit is extended. It sends back dial tone to the calling subscriber if it is a first selector. It elevates the shaft and wipers in response to dial pulses and rotates them automatically to connect with an idle trunk in the selected bank level. It provides a busy signal to the calling subscriber when all trunks in the desired group are busy. The selector is returned to normal, when the calling subscriber places his handset on its support, by functioning of the control circuit and the release magnet of the selector.

The connector is a two-motion switch, similar to the selector, and, regardless of the size of the office, it is always employed as the final unit of step-by-step switch trains.

This switch operates in response to the last two digits dialed of the called number (directory listing). The first of these two digits is the "tens" and the last one the "units" digit. The only exceptions to this general principle are the 200-line connector and the frequency or code-selecting, party-line connectors, where a digit preceding or succeeding the "tens" and "units" digits is dialed for line group or ringing selection.

Figure 13 shows a schematic diagram of the connector circuit. The connector has a group of control relays mounted on a base, which also supports the shaft and ratchet

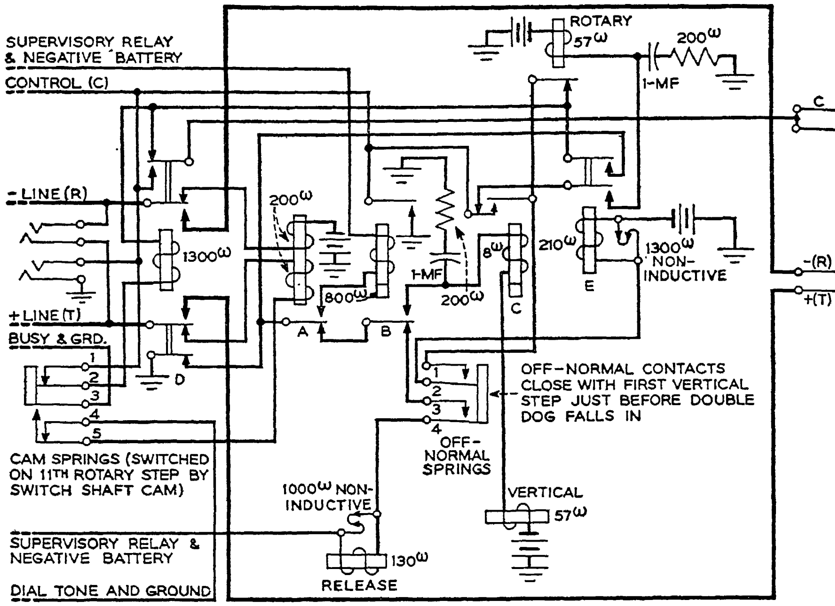


Fig. 12. Selector Circuit—Strowger System (Courtesy Automatic Electric Co.)

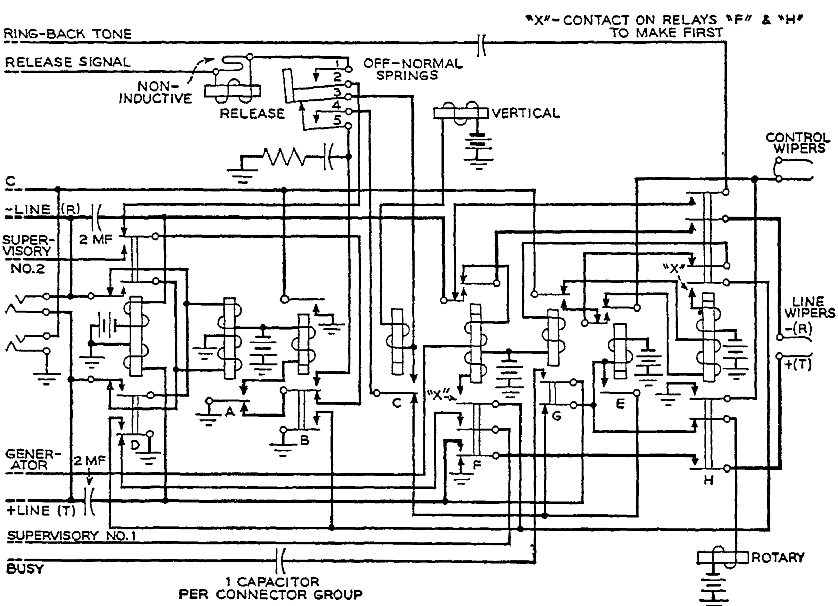


Fig. 13. Connector Circuit—Strowger System (Courtesy Automatic Electric Co.)

mechanism assembly for raising and rotating the shaft, step by step. The lower part of the shaft carries two sets of wipers, the *control* (upper) and *line* (lower), which engage respective semicircular banks of contacts of 100 sets each (10 levels and 10 sets per level). The vertical and rotary stepping magnets and the release magnet are mounted within the switch frame. One subscriber line is connected to each set of lower line bank contacts.

The pulses of the first digit received (except as mentioned above) step the shaft and wipers vertically as many levels as there are pulses. The pulses of the last digit received step the shaft and wipers horizontally, in accordance with the number of pulses received. This desired position of the shaft assembly is held by two movable detents, termed the "double-dog," and a stationary dog.

The release of the shaft is under the control of the release magnet, which, in operating, disengages the dogs, allowing the shaft assembly to return to normal.

The director system was originated and developed by Automatic Electric Co. for large and complex trunking networks of the Strowger system. Though not used in the United States to date, it has extensive application in Great Britain, particularly in the London metropolitan area. It is being considered in the Los Angeles area for certain S × S offices to meet extended service and automatic toll ticketing problems. The director is expected to play an important part in nationwide toll dialing, requiring register-sender equipment.

The director, itself, consists of standard A. E. Co. relays and switches, which store the subscriber dial pulses and perform various other functions. A director for simple functions occupies the space of two regular switches, but, being very flexible in design, it may be of various sizes for specific needs. Wherever used, this unit usually effects savings in selectors, repeaters, and floor space. It can be added to existing S × S equipments as desired.

In operation, an idle director is selected by a director finder and attached to a line finder-first selector trunk as soon as the line finder seizes the calling line. The director then functions as an "electrical brain" to:

1. Record the number (pulses) dialed by the subscriber.
2. Analyze the office code digits received and immediately determine the best routing for the call.
3. Substitute, if necessary, other routing digits, which may differ entirely from the received digits. This is known as translation, which permits automatically selected alternative routings with resulting trunk savings.
4. Send out pulses, corresponding to the routing code, which operate switches, as though operated by the subscriber dial.
5. Store the remaining digits, dialed by the subscriber, usually without translation, and send out corresponding pulses just after the routing code is sent.
6. Detach itself from the connection upon completion of operation 5 and await the next call.

The toll switch train, consisting of a toll transmission selector, a toll intermediate selector, and a combination toll and local connector, is designed to complete toll calls directly from the toll board to the subscriber. The *toll transmission selector* provides increased talking current to the called subscriber telephone through repeating coil windings and has increased capacity in the talking circuit. It also provides for complete supervision of the connection at the toll board, and it repeats the dial pulses from the toll operator dial through to the toll intermediate selector and the toll connector. It has a 400-point bank, required for additional functions. The *toll intermediate selector* is similar to the regular selector except that it also has a 400-point bank. The *combination connector* is not used for local calls unless all the regular connectors are in use.

Main distributing frames (MDF) and intermediate distributing frames (IDF) provide a means for properly terminating the outside cables, which carry subscriber lines and various types of trunks and toll circuits, and equipment within the office, and cross-connecting these various circuits and equipments as required.

Multipling of trunks and of switching equipment is so arranged as to provide maximum access to switches from subscriber lines and from switches to lines, and efficient operation of trunking and other equipment.

Figure 14 shows a schematic diagram of a trunk layout for a 100,000-line multioffice exchange. It will be noted that secondary line switches are employed between the primary line switches and first selectors and between the first selectors and impulse repeaters, the object being to concentrate the traffic loads in these sections and reduce to a minimum the number of first selectors and repeaters required. All manual board services are centralized at one office for efficient operation. A special switch train is provided to reduce wrong numbers which may result from careless removal of a handset or accidental jiggling of the cradle plunger switch. This equipment is not always warranted.

The *U* suboffice represents a telephone center of light load, located rather remotely from the other offices. Traffic does not justify a separate group of trunks to each of the other offices, so that the office is made tributary to office *U*. To reach a subscriber in the

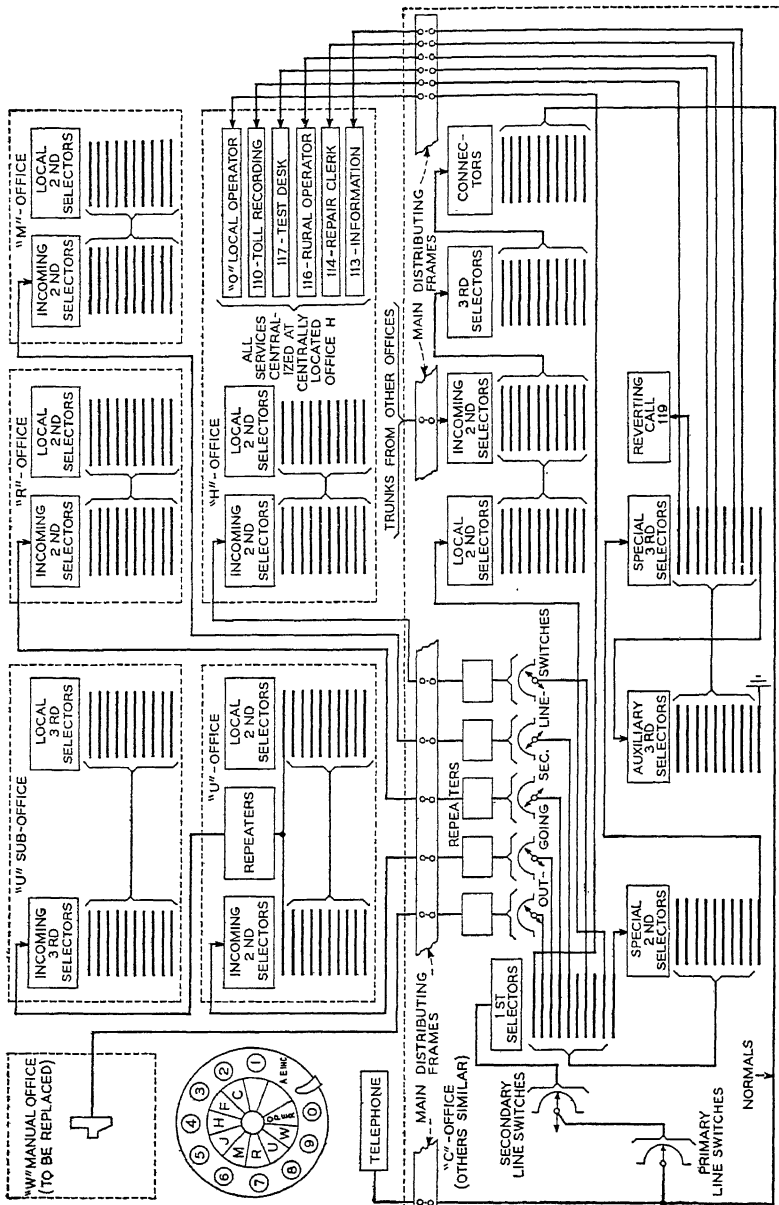


Fig. 14. Trunk Layout for a 100,000 Line Multioffice Exchange—Strowger System (Courtesy Automatic Electric Co.)

office, *U*0 is dialed, the first incoming switch in the *U* suboffice being a third selector. Outgoing calls from this office pass through the main *U* office.

Multioffice trunking and other switching in the Strowger system is accomplished, using in combinations the types of step-by-step switches and repeaters described above and, in addition, many other auxiliary circuits and devices which limited space will not permit discussing here.

THE PANEL DIAL SYSTEM is a radical change, both in mechanical and electrical characteristics, from the Strowger system and is employed generally in the largest metropolitan areas. Great flexibility of operation is permitted by this system as trunk groups and their sizes are for all practical purposes arbitrary. The time of establishing a connection is not dependent on the step-by-step dialing process of the subscriber, since, after all the dialing pulses are received from the subscriber by an intricate mechanism, the connection is rapidly routed to the called line, under control of this same mechanism, with great speed. The equipment as a whole is necessarily complex.

The principal mechanisms in the system are:

1. The panel-type selector with its banks of line or trunk terminals, which are selected by vertically moving brushes mounted on brass rods.
2. The sequence switch, which has a number of insulating disks mounted on a shaft. Each disk has a metal stamping on each side, and the entire assembly of disks turns a step at a time, as directed by the control circuit, with two brushes bearing on each stamping. By this means a large number of circuit arrangements may be made as required to establish the talking connection at the proper time, as determined by the control circuit.
3. The decoder sender is the "brains" of the system. It registers and stores the dial pulses from the subscriber dial by means of a dial pulse register circuit: it decodes the numerical digits dialed into a non-numerical selection scheme by choosing the proper route relay. The route relay controls the selection of an interoffice trunk. The sender then takes over control, sending out pulses to actuate a train of selectors which establish connection with the called line. The decoder sender is then released for other calls.

The panel-type selector, from which the system derives its name, performs the same function in this system that the line finder and selector switches perform in the Strowger system. The panel-type selector frame consists of several panel multiple banks of subscriber line or trunk terminals, over which the selector brushes slide vertically, and mechanism for moving and controlling the selector motion, as shown in Fig. 15.

The panel multiple bank consists of horizontally projecting terminals, arranged in vertically positioned, rectangular panels or banks, as shown in Fig. 16. Each terminal in the panel consists of a flat brass strip extending horizontally through the panel and having 30 projections on each side of the panel and also a soldering lug at each end of the strip for wiring. A set of three of these strips, mounted one above the other and insulated from each other with impregnated paper, constitutes the tip, ring, and sleeve terminals of one line. Thus, each line appears horizontally across the face of the panel 30 times, or 60 times for both faces. The number of lines or trunks provided in the banks varies in accordance with requirements, present practice with respect to subscriber lines being to provide 40 lines per panel and 10 panels per frame.

A selector is placed opposite each three vertical rows of line terminals; it consists of a hollow vertical brass tube on which are fastened ten sets of spring brushes, one set of brushes for each terminal bank. Each brush has three contacts, normally held apart by an insulator so as not to touch the lugs or terminals on the terminal panels. Each brush contact is connected in multiple with the corresponding contact of the other nine sets of brushes on the same brass tube or selector, so that any selector may reach any one of the 400 lines in the frame by tripping the proper brush, and the total selector movement will not exceed one of the ten banks of terminals. In practice, each bank of 40 lines is divided vertically into two identical banks of the same lines, but with the line numbering in reverse order (bottom to top and top to bottom of the bank). This arrangement of line numbering still further reduces the travel of the selector brushes to reach a particular line and reduces operating time of the selectors. A trip rod is provided with trip levers for each bank, so that the brush which is selected to make contact with the calling line is tripped as it starts upward in the calling line bank, the insulator between the brush contacts is withdrawn, and these contacts make the connection with the calling line terminals when they are reached.

The multiple wiring between the brushes is contained inside the hollow brass tube of the selector and terminates in another set of brushes at the top of the frame, which slide on bars in the commutator panel and control the selector movement.

The control mechanism of the selector is located at the bottom of the frame. There are two horizontal cork-covered rolls, extending the width of the frame and driven in opposite directions by suitable gears and a motor. Attached to the lower end of each selector tube is a flat bronze strip or rack, which is close to, but, in the idle position, not touching, the continuously revolving cork rolls. An electromagnetically operated clutch is mounted on the selector frame in front of each rack. Energization of either the up or down magnet presses the rack against the up or down cork roll, causing the selector to move up or down as desired. A pawl (not shown) engages in the horizontal slits of the rack and prevents the selector from slipping down after it has been elevated to the proper level. The selector returns to its normal position when released by the trip magnet. Figure 16 shows the control mechanism.

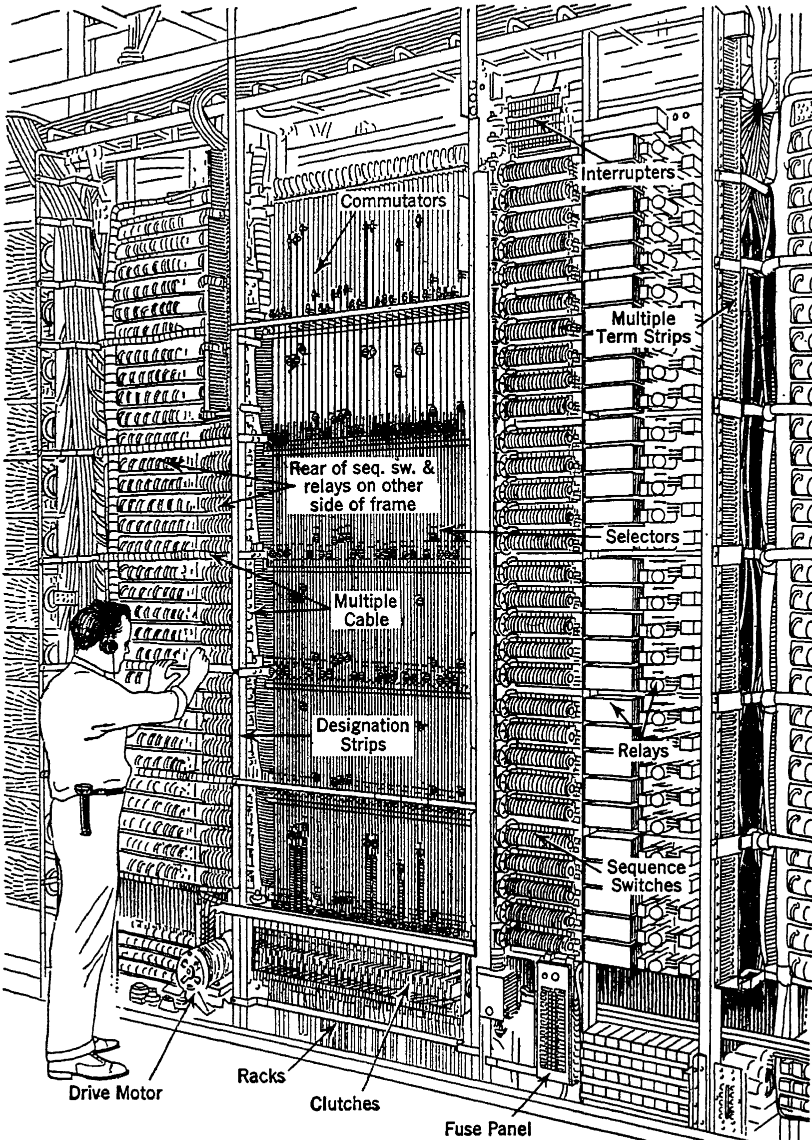
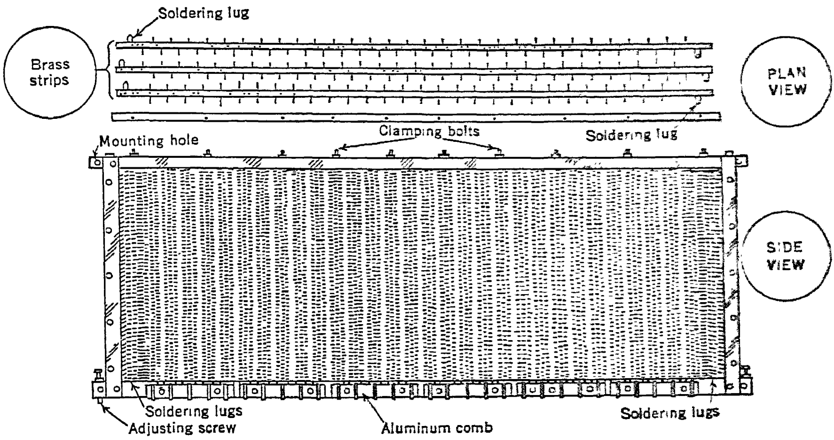
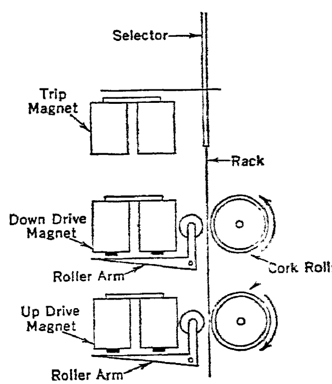


FIG. 15. General View of Panel Frame—Panel Dial System (Courtesy Bell System)

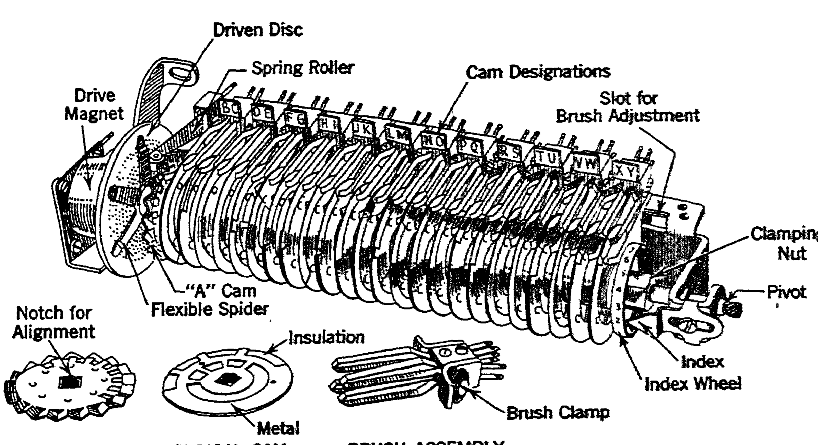
Control of the up-and-down movement of the selector is obtained either from impulses coming from the sender or from connections made on the commutator at the top of the frame. Panel selector equipment may vary in construction details and wiring, depending on its assigned function in the system.



Panel Multiple Bank



Selector Control Mechanism



Sequence Switch

FIG. 16. Panel Frame Units, Showing Details—Panel Dial System (Courtesy Bell System)

The sequence switch is vital to panel operation. The control circuits for panel-type selectors are necessarily complex and must be set up in sequence to cause the various operations for establishing a connection to take place at exactly the right time. This switch (Fig. 16) has 24 disks or cams mounted permanently on a shaft, and each disk consists of insulation with a specially shaped metal stamping attached to each side of the disk. Two brushes bear on each side of each disk. The metal stampings are of different shapes, so that as the disks turn each brush may rest on the stamping or on the insulation, in accordance with the position of the switch. This provides a means of establishing connections between brushes or opening and closing circuits in various combinations, as desired. A fluted metal disk, on which a spring roller rides, is mounted at the end of the shaft, so that the shaft may be revolved and held in any one of 18 positions. An electromagnet, when energized, pulls this disk against the edge of another continuously revolving disk and thus causes the shaft to turn one step at a time for each energization of the electromagnet.

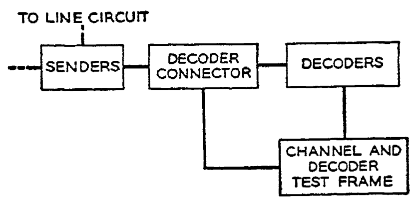


FIG. 17. Interconnection of Decoder Sender Units—Panel Dial System (Courtesy Bell System)

One sequence switch is associated with each selector and is located on the right of the terminal bank, as shown in Fig. 15.

The decoder sender records the pulses dialed by the subscriber, translates and decodes the first group of pulses (office code), resulting in selection of the proper central office trunk or trunks, and finally controls the brush selection and functioning of the various selectors and the sequence switch. The major units consist of a sender, decoder connector, and decoder, as shown in Fig. 17.

The dial pulse register circuit of the sender consists of a group of relays so connected as to record the dial pulses and subsequently control the operation of the decoder and selectors in establishing the connection. Figure 18 shows the portion of the dial pulse register circuit which registers the first three series of pulses comprising the office code. The connection of the sender to the subscriber's line operates the *AC* relay, so that the *P* group of relays is connected to the *A* group. The *P* group counts the pulses as received from the subscriber dial. The release and reoperation of the *L* relay, under control of the dial, operate the *P* relays. The *A* recording relays are operated by the action of the *P* relays in such manner that the sum of their numbers is equal to the figure dialed. Thus, if the first letter of the office name is *S*, causing seven pulses to be sent, relays *A*-2 and *A*-5 will operate. When the first figure has been recorded, the *RA*-1 relay supplies ground to the *A* recording relays, locking them, and releases the *P* relays. The *P* relays are then connected to the *B* recording relays, which are ready to record the next figure dialed. This operation is similar for each of the remaining figures dialed by the subscriber.

There are eight groups of recording relays, three for the office code and five for numerical digits. The last five groups are not shown in Fig. 18, but they are similar to those shown.

When three digits have been dialed, relay *CL* operates (relay *Az* operates in place of relay *CL* if the first figure dialed is zero). The operation of either of these relays indicates that the sender is ready for translation. The sender is now automatically connected to a decoder, through the decoder connector, by connection of the *A*, *B*, and *C* groups of leads shown in Fig. 18 to the corresponding groups of leads shown in Fig. 19.

The decoder register relay groups, *A*, *B*, and *C*, shown in Fig. 19, then register the hundreds, tens, and units digits of the office code, stored in the sender. The setting of the *A* register relays causes one of the multicontact relays, *H*-2 to *H*-9, to operate; the setting of the *B* relays causes operation of the proper *T* relay, and the setting of the *C* relays grounds one of the 10 leads shown at the right in Fig. 19.

Each *H* relay has 100 armatures and contacts as indicated in Fig. 19. The 800 contacts are connected to terminal strip punchings, called code points. The operation of an *H* relay connects the proper hundred code points to the hundred contacts of the *T* relays (ten to each relay). The operation of the proper *T* relay connects ten code points to the ten leads from the *C* relays, and the setting of the *C* relays connects ground to one code point.

The code points are cross-connected to route relays (Fig. 19), one of which is provided for each possible path, which a call may take. Since a particular route relay may be cross-connected to any code point, so that the selection of a particular route relay is numerical, the selection operation from this point on need not be on the decimal basis as indicated above.

A particular route relay is so designed that it requires certain operations of the district and office selectors independently of the code point to which it is connected, and thus of the number dialed in securing it. If the position of a group of trunks is to be changed on the selector, the route relay must be changed accordingly, but no further change in the system need be made. Only one route relay is required in each sender for each outgoing

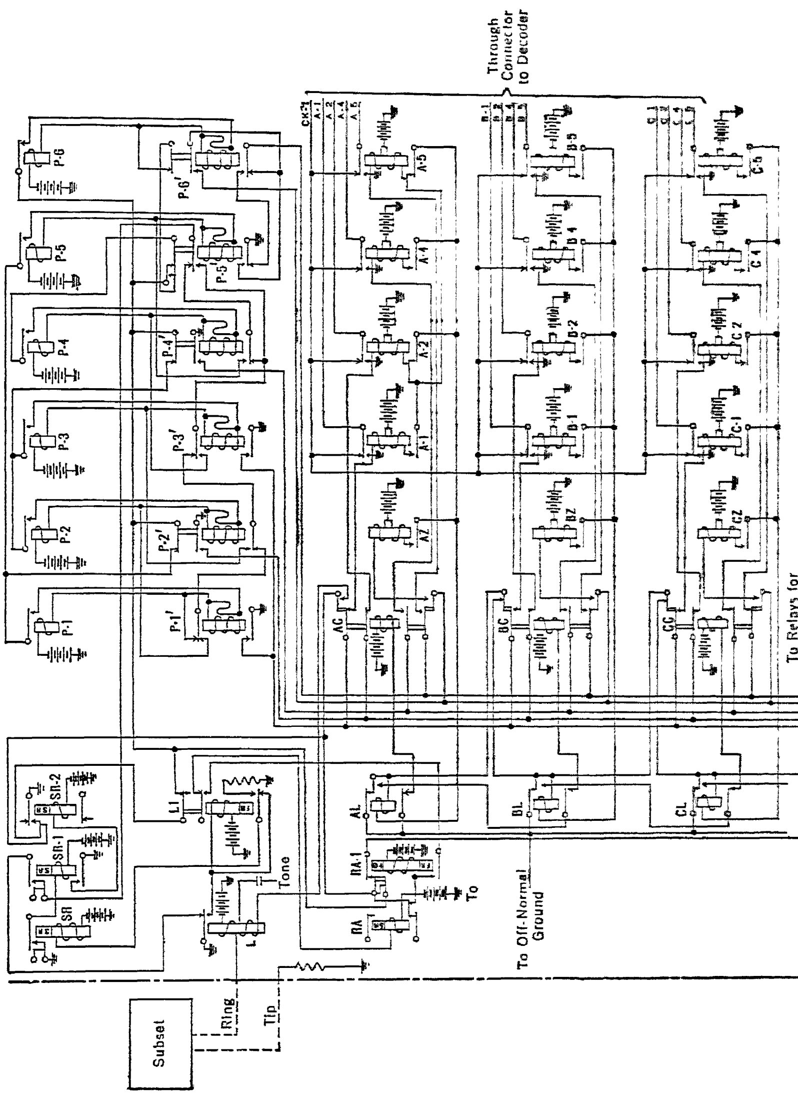


FIG. 18. Dial Pulse Register Circuit of Sender—Panel Dial System (Courtesy Bell System)

trunk group from the district and office multiple, that is, one for each dialing combination used in the particular exchange. The route relay, through a combination of registering relays, determines the brushes to be tripped and the movements of the selectors which choose the proper office. Once the proper route relay is selected, the remaining operation of obtaining the proper office is straightforward. All vacant code points are strapped together and connected to a single route relay, which will cause the sender to complete the call to an operator, who will explain to the subscriber that an incorrect number has been dialed.

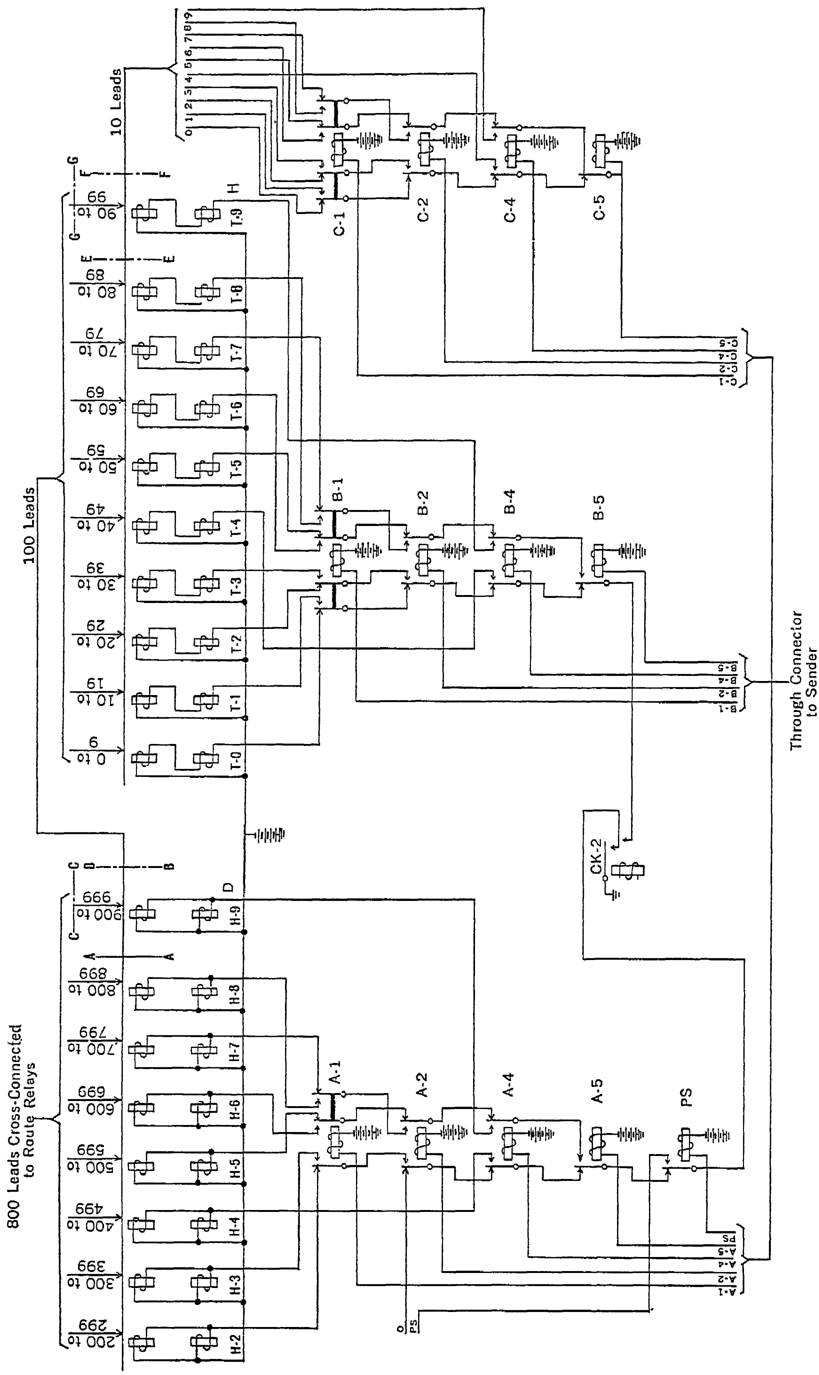


FIG. 19. Decoder Register Relay Circuit of Decoder—Panel Dial System (Courtesy Bell System)

When the decoder has completed translation, or if translation cannot be completed owing to trouble, the decoder signals the sender, which releases the decoder. Thus, the decoder is needed only during the time the office trunk is being chosen, but the sender is needed in the connection until the called subscriber line is connected. More senders will be required than decoders, the usual ratio being about 400 senders to not over 10 decoders, although this ratio varies with the traffic load.

IN PANEL DIAL OPERATION each subscriber line terminates on a line finder frame for handling outgoing calls, and a final frame for handling incoming calls, as shown in Fig. 20. When a subscriber lifts his handset to call, direct current from the central-office battery flows through his line and operates the associated line relay, at the same time making his line busy to any selector on the final frame. The operation of the line relay causes an idle line finder selector associated with the bank of terminals in which his line appears to move upward until it finds the calling line; this relay also causes the trip rod in that bank to trip the proper brush of the moving selector, after which the trip rod is immediately reset.

Simultaneously with the motion of the selector, the panel link circuit is selecting the district selector associated with the line finder selector, and also an idle sender. The panel link frame is also similar to the panel type selector frame, but with different details to suit its functions. The frame is divided into two equal parts vertically with the sender selectors on the left and the district selectors on the right. Each district selector is wired directly to its corresponding sender selector.

In practice, the district selector when released from a call selects the next idle line finder, while the sender selector remains where it is unless near the top of the bank, in which case it returns to normal position.

When the line finder reaches the calling line and the sender selector has found an idle sender, dial tone is sent back to the calling subscriber.

The action of the sender is described in detail. The dial pulses are stored in the dial pulse register circuit, the first three digits serving to connect the proper route relay to the sender and thus set the proper registering relays in the sender, at which point the decoder is disconnected from the circuit.

The sender is then prepared to cause the district selector to choose an interoffice trunk to an incoming frame. The district selectors are typical panel-type selectors, except that every eleventh terminal and an additional one between the two groups of five at the top of each bank is arranged as an overflow terminal to stop the motion of the selector when all ten trunks below it are busy. The trunks can be used in larger groups than multiples of 10 by making the intermediate overflow terminals busy, in which case the selector will pass over them, or groups of five trunks are available at the top of each bank.

If 450 trunks are not enough to provide the necessary trunks to all the offices in the exchange area, a group of office selectors is employed to care for all the trunks outgoing from the office. The calls are then routed by the district selector to the proper office selector and thence to the incoming selector of the called office.

The movements of the selectors on both district and office frames are guided by a reverse control method. As the selector is driven upward by the cork roll, it sends back one pulse to the sender each time a brush moving on the commutator (at the top of the frame) makes contact. Commutators for the various types of selectors are shown in Fig. 21. The pulses are counted by the sender, and, when the number of pulses indicates to the sender that the selector has moved to the proper position, the sender opens the up-drive magnet circuit, stopping the selector motion.

The first selection on either the district or office frame involves the tripping of the proper brush. If, as shown in Fig. 20, the desired trunk appears on the fourth panel from the

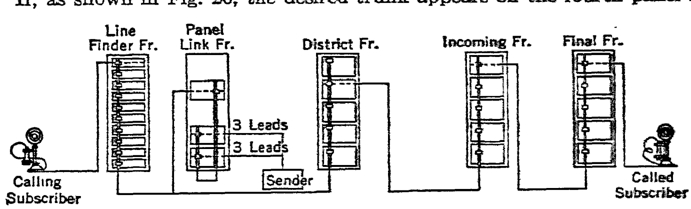


Fig. 20. Typical Panel-type Connection—Panel Dial System (Courtesy Bell System)

bottom, the district selector will first move up a distance of four segments on the *A* bar of the commutator (Fig. 2 in Fig. 21). At that point the selector will be stopped, the sequence switch will be turned two steps (from position 4 to 6 in Table 1), changing the fundamental circuit to prepare for the next operation, and the trip rod will be rotated into position to trip the fourth brush. The selector is now started upward again by the sender, and sends back a pulse for each bar on the *B* bar of the commutator, each of which represents a distance of ten trunks. Thus, the upward movement of the selector to the proper trunk group is controlled by the number of pulses sent to the sender from the *B*

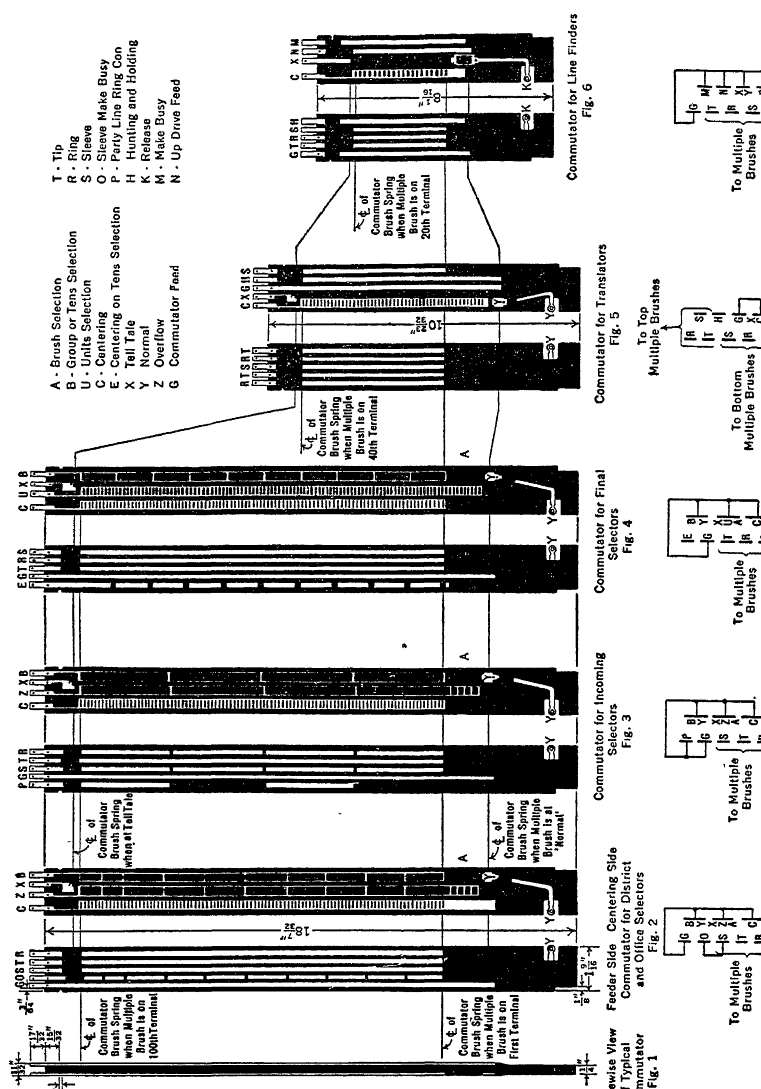


FIG. 21. Commutator Brush Connections for Selectors, Translators, and Line Finders—Panel Dial System (Courtesy Bell System)

bar. When the selector reaches the proper group, the sequence switch is again turned two steps (position 6 to 8 in Table 1) and the selector moves up slowly, testing each trunk in the group to find an idle one—a process known as hunting.

The hunting process is common in panel operation. The connections are shown in Fig. 22, all connections marked 8 being closed, since the sequence switch is on position 8 (Table 1). If the *L* relay is operated, the up-drive clutch is energized by current flowing to ground through the *L* relay left-hand contact, and the selector moves up. As long as the sleeve contact of the selector brush is grounded, the L relay is operated through contact 7 1/2/8. Since the sleeve contacts of busy trunks are grounded and those of idle trunks are not grounded, the selector will move a small distance beyond the last busy trunk. However, even here the L relay is grounded through the commutator brush and the G bar of the commutator. The reason for this slight extra movement is to allow the holding pawl on the selector rack to engage the correct slot.

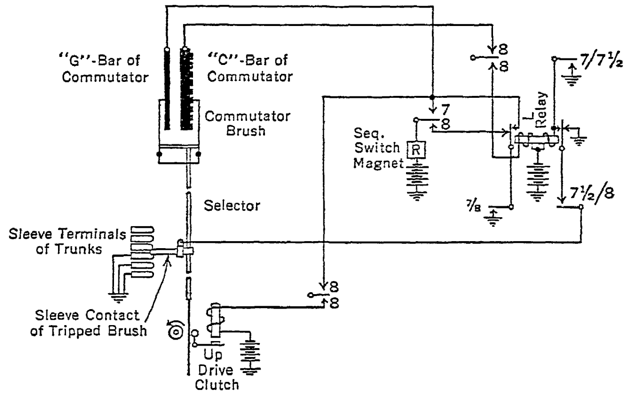


FIG. 22. Selector Circuit for Hunting Idle Trunk—Panel Dial System (Courtesy Bell System)

When an idle trunk is found, the sequence switch turns to position 10, until the connection is completed, the changes necessary in the control circuits being made by other sequence switches located on the incoming and final selector frames.

Table 1. Operations of District Sequence Switch

Position	Corresponding Circuit Condition	Position	Corresponding Circuit Condition
1	Normal	10	Selection of brushes, groups, etc., beyond the district selector
2	Selecting an idle sender		
3	Waiting for sender	11	Waiting for sender
4	Selecting brush	12	Talking (non-loaded trunk)
5	Waiting for sender	13	Talking (medium-loaded trunk)
6	Selecting group	14	Waiting for operator to answer
7	Waiting for relays	15	Talking to operator
8	Hunting idle trunk	16	All trunks busy
9	Waiting for sender	17	Operating message register
		18	Returning apparatus to normal

These selectors are also of the panel selector type, differing only in details. The final selector frame has capacity for 500 lines, thus requiring 20 final frames for a 10,000 subscriber line office. The incoming selector frame must then hold 20 groups of trunks, which can consist of 24 trunks and an overflow terminal. This arrangement, which is employed in all offices, is shown in Fig. 23. The sender circuit is so arranged that either 0 or 1 in the thousands digit causes brush 0 to trip, while the combination of the first two numbers determines the proper group for the selector to hunt in. When an idle trunk to the final selector is found, the second or hundreds digit determines which brush will be tripped, while the tens digit determines in which group the called line appears. Final selectors are directed rapidly to the proper group of ten; when the proper group is reached, the clutch on the up-drive roll is released by a change in the setting of the sequence switch and 1/4 usual speed is substituted as the motive power. The final selector tests the line on which it stops, and if it finds the line busy it returns to normal and sends back a busy tone to the calling subscriber.

If the called subscriber has a private branch exchange (PBX) with several trunks, the final selector will hunt for an idle line if the called number is busy exactly as it hunts for an idle interoffice trunk.

Many special provisions must be made in any installation for party lines, message rate service, rapid testing, and other special conditions that arise, but such arrangements will not be considered here.

THE CROSSBAR SYSTEM is the latest Bell System development in mechanical switching. It differs radically from the Strowger (step-by-step) and the panel dial systems in construction, operation, and control, but it is so designed that it functions satisfactorily with all other types of switching, whether mechanical or manual. The crossbar system offers important improvements in switching, both in operation and maintenance, over the step-by-step and panel systems, but it does not necessarily replace existing installations or additions to these older systems.

Only the most important functions of the crossbar system can be discussed in this handbook, owing to space limitations.

The crossbar system has two outstanding features, *the crossbar switch* which is used for all major switching operations, and *the marker system* of control which is used in establishing all connections throughout the crossbar office. The apparatus consists principally of crossbar switches of the relay type and multi-contact and other type relays generally employed in telephone systems. The switching circuits are wired to the contacting springs of the switches, and the circuit closures are made when the contacts are pressed together by operation of the electromagnets.

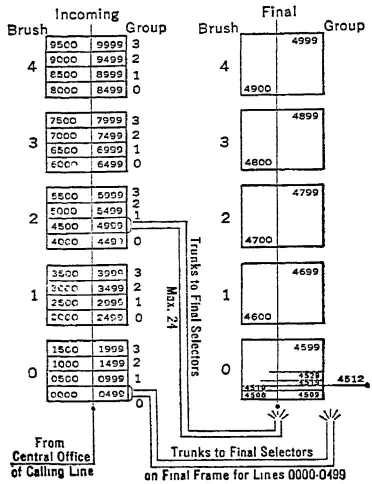


FIG. 23. Trunk and Subscriber Multiple on Incoming and Final Selectors—Panel Dial System (Courtesy Bell System)

The use of relay-type apparatus economically permits having twin contacts of precious metal throughout, insuring reliable operation for the low values of speech and signaling currents inherent to telephone systems. The very short mechanical movements and small operating time intervals required in crossbar switch operation permit a reduction in control equipment over the slower-moving, older-type systems, thus resulting in the use of large switch and relay assemblies on unit-type frames, factory wired and tested. In the design of the switching frames and associated control circuits, it has been possible to standardize a relatively few types of equipment units, thus simplifying manufacture, merchandising, and operating company engineering.

The marker, composed mainly of relays by means of which it controls switching operations, has many advantages, one of the most important being the completion of its complex functional processes in establishing a call in less than one second. The markers are connected momentarily, by means of multicontact relays, to various switching units of the office to guide the completion of calls through the crossbar switches. The marker system provides for an attempt automatically to establish a call over alternative paths when all the normal routings are busy or trouble is encountered; the marker will detect, record the location and nature of such troubles, and indicate their presence to an attendant by operating an alarm. The marker design also facilitates the introduction of new service features and operating changes as desired, since the main control features of the entire system are incorporated in a small number of markers.

The crossbar switch, which gives the system its name, is the basic switching unit of the system. Figure 24 shows a front view of a 200-point switch. Fundamentally, this switch contains (1) 10 separate horizontal circuit paths, (2) 20 separate vertical circuit paths, and (3) an electromagnet for each horizontal and each vertical path, so that the operation of any horizontal and any vertical magnet in sequence will connect a horizontal and a vertical path together at one of the 200 cross-points. The 10 horizontal paths are controlled by 5 horizontal bars each actuated by a selecting magnet, and the 20 vertical paths are controlled by 20 vertical bars each actuated by a holding magnet. Any set of contacts may be closed by first operating the selecting magnet corresponding to the horizontal row in which the contacts are located, and then by operating the holding magnet associated with the particular row of vertical contacts. The holding magnet will hold the contacts closed until the connection is released, but after it is energized the selecting magnet returns to normal until called upon to operate on another call. Thus, 10 sets of contacts may be made at one time through the switch, one for each horizontal path.

Figure 25 shows in detail a portion of the selecting mechanism of the crossbar switch. The 10 sets of contacts in each vertical row are associated with the vertical or holding bar of that row. Each horizontal or selecting bar has 20 flexible wire selecting fingers, mounted at right angles to the bar, one finger for each vertical row of contacts. When a selecting bar is rotated through a small arc by its magnet, the selecting fingers move up or down into position, depending upon the direction of rotation of the bar. When a holding bar in a

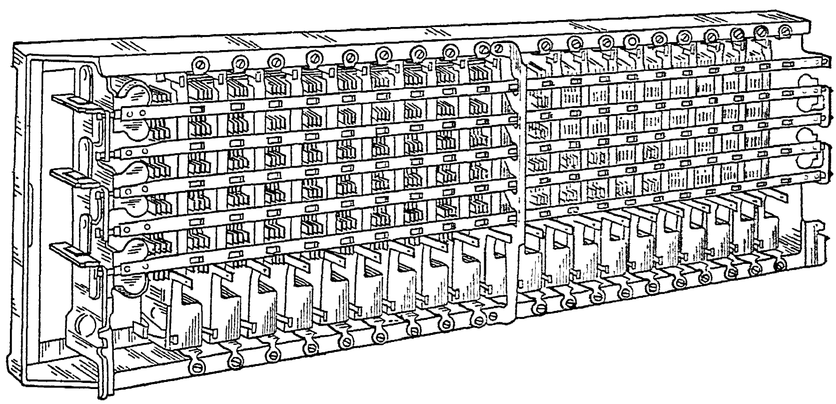


FIG. 24. Crossbar Switch—Front View—Crossbar System (Courtesy Bell System)

particular vertical row is operated by its magnet, it will bear against the particular selecting finger which has been moved into position in its row, and the finger will be pressed against the operating spring horizontally in line with it. Thus, the operating spring will be pressed against the contact multiple (fixed contact spring) and the circuit path will be closed at that point.

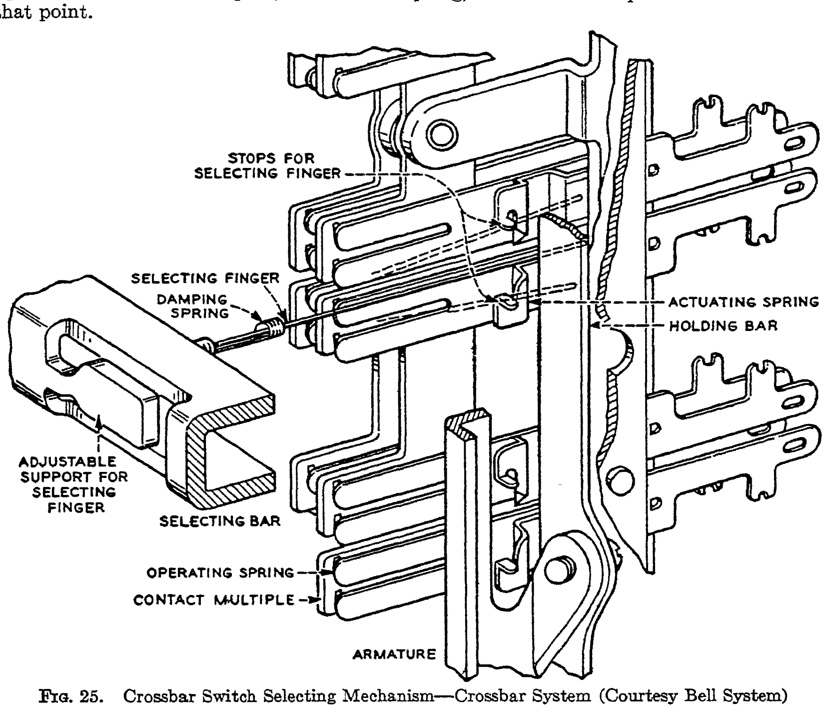


FIG. 25. Crossbar Switch Selecting Mechanism—Crossbar System (Courtesy Bell System)

The selecting bar and all its fingers except the one being held against the operating spring will be released as soon as the holding bar operates. When the connection is released, the holding bar returns to normal and the held finger returns to its idle position midway between, but slightly to the right of, two sets of operating springs. At those points along an operated vertical bar where the fingers are not operated by the horizontal bar, the fingers are pushed in between each two sets of operating springs and thus do not bear against these springs. Only one finger is operated at one time for a given vertical bar.

The selection operation is performed by five horizontal bars, each of which will select one of two horizontal rows of operating springs, depending on the direction of rotation of the bar and consequently the selecting fingers. Figure 24 shows two magnets at the end of each horizontal bar, one causing the fingers to move upward and the other causing them to move downward. After release, the horizontal bar is restored to its mid or idle position by centering springs at the ends of the bars and adjacent to the magnets.

The vertical units of the crossbar switch each have 10 sets of operating springs in vertical rows with one vertical or holding magnet at the bottom of the row which actuates one vertical bar or armature extending the height of the 10 sets of contacts. Each set of contacts may consist of three, four, five, or other numbers of pairs of springs in horizontal spring pile-up or assembly, depending on circuit requirements. Figure 24 shows four pairs of springs per set of contacts. One twin contact spring of each pair is stationary and designated contact multiple; the mate or operating spring of the pair is pressed against the contact multiple, when operated. This contact multiple spring is made of one piece of metal, insulated from the mounting and extending the full length of a vertical row of contacts. Wiring lugs are formed at the lower end of these vertical metal pieces, facing the rear, to which are wired the lines or trunks of the vertical circuit paths. On the front and at the lower end of these pieces projections are provided for testing purposes. The mate or operating springs extend to the rear of the switch, where wiring lugs are provided and may be strapped horizontally to the corresponding springs of adjoining vertical units, thus extending the horizontal paths across the switch verticals, or to adjoining switches.

The switch may have "off normal" contact spring assemblies, if required, associated with each selecting magnet which operates them to perform various circuit functions.

The 200-point crossbar switch is 9 1/4 in. high and 30 1/2 in. long. A 100-point switch with 10 verticals is also available.

The multicontact relay used in the crossbar system and shown in Fig. 26 resembles in design the vertical unit of the crossbar switch. The relay is provided with 30, 40, 50, or

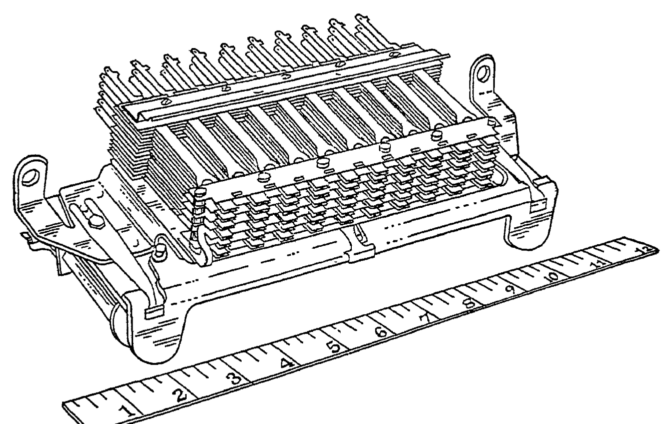


FIG. 26. Multicontact Relay—Crossbar System (Courtesy Bell System)

60 sets of individually insulated contacts, normally open, but closed when the relay magnets are operated. Each relay has two separate magnets, armatures, and associated groups of springs. By operating the magnets independently, the unit can be operated as two separate relays, each closing 15, 20, 25, or 30 sets of contacts, or if both magnets are energized in parallel the full number of contacts of the unit are closed. All contact springs have twin contact surfaces of solid bars of precious metal because of heavy duty requirements. This relay is used mainly in the common connector circuits, where a large number of leads must be connected simultaneously to a common circuit.

New and improved general-purpose small relays of the U and Y types are used in the crossbar system. These relays permit the use of up to 24 springs in one assembly, providing for various combinations of transfer or simple make and break contacts. The springs are equipped with twin contacts of various contact metals, depending on the characteristics of the circuits controlled by them. The Y relay has a slow release action obtained by copper or aluminum sleeves over the relay core.

CROSSBAR OPERATION may be more readily understood from the block diagram of the functional arrangement of the principal equipment units of the system which are involved in a connection between two subscribers, as shown in Fig. 27. The three main

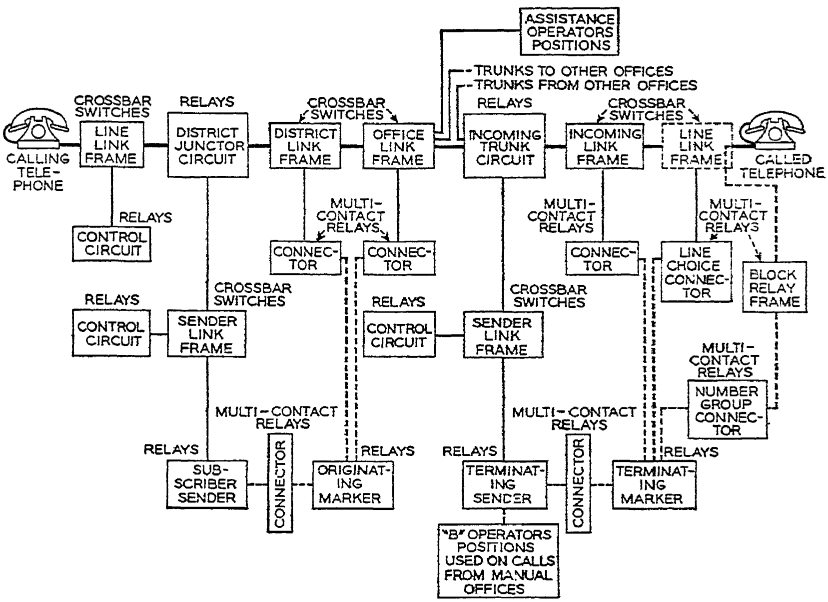


FIG. 27. Functional Arrangement of Equipment Units—Crossbar System (Courtesy Bell System)

types of equipment units are: (1) the district junctors and incoming trunks, which supply battery to the transmission and supervisory circuits; (2) the crossbar switch frames; (3) the common control circuits and the senders and markers.

The district junctor and incoming trunk circuits are composed mainly of small relays. The district junctor furnishes talking battery for the calling subscriber and supervises the originating end of the connection. The incoming trunk circuit controls the ringing of the called subscriber bell, furnishes talking battery for the called line, and supervises the terminating end of the connection.

The switch frames, consisting principally of crossbar switches, provide the means for switching between the subscriber lines, district junctors, and incoming trunks, and also for switching these district junctors and trunks to senders.

The senders consist chiefly of small relays, and their function in crossbar is comparable to that of an operator in manual operation. The subscriber sender registers the called number from the subscriber dial pulses and transmits the necessary information to the markers, to the terminating sender, and to the manual operators (in manual offices) for completing connections to the called line. This sender also controls operation of the selectors in distant panel offices. The terminating sender receives the numerical digits of the called number from the subscriber sender of any panel or crossbar office and transmits the required information to the terminating marker for setting up connections to the called line.

The markers are the most important control circuits in the crossbar system. They comprise both small and multicontact relays and are of two types, one for originating and one for terminating traffic. Since their operating time is less than 1 sec, only three or four markers are needed in the average size office.

The originating marker determines the proper trunk routes to the called office. It has access to all outgoing trunk circuits and all crossbar switch frames used in establishing connections to the called office trunks. The marker records pulse information, tests the trunk group for an idle trunk to the called office, tests for and marks or reserves an idle path through the switch frames, and finally operates the proper crossbar switches to establish a path from the calling subscriber line to the outgoing trunk circuit. The lowest-numbered available paths are always selected in order to reduce selection time and increase operating efficiency.

Trunk selection is made by the marker through route relays, of which there is one for each called office routing. This relay is so wired as to direct the marker to the called office trunk group and indicate the number of trunks in the group; also to indicate the office link switch frame on which the trunk group appears and the type of called office, which is also indicated to the sender. Route relays may be assigned to any office trunk group, and other changes may be made, as required.

The terminating marker performs similar functions in the terminating office, establishing a path between the incoming trunk circuit and the called subscriber line. This marker has access to all the subscriber lines terminating in the office and to all crossbar switch frames used for connecting to subscriber lines. It records pulse information, tests the called line to determine whether it is idle, tests for and marks an idle path through the switch frames, and finally operates the proper crossbar switch magnets to establish connection to the called line.

The marker, in testing and connecting the called line, employs a marker group connector circuit and block relay frame in which the called line appears. Each subscriber line has three test terminals on the block relay frame, and a number group connector will usually have access to the test terminals of several hundred lines. The marker determines from these test terminals whether the called line is busy and the identification of the proper line link frame and horizontal group of line links which have access to the called line; also the type of ringing required is determined from circuit conditions on the test terminal.

There are also *common control circuits*, associated with the line link and sender link frames, for controlling the operation of the switches on these frames. In addition, there are *common connector circuits*, composed mainly of multicontact relays, which are used to connect the markers (1) to their respective senders, (2) to their associated switch frames, and (3) to the subscriber line test terminals.

The line link frames, although shown separately in Fig. 27, are, in a given office, used for both originating and terminating traffic. After the talking connection has been established between subscribers (see Fig. 28), all the common control circuits, including senders, markers, connectors, line link control circuits, sender link frames, and their associated

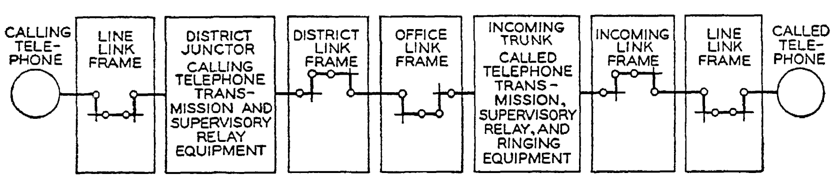


FIG. 28. Completed Talking Connection—Crossbar System (Courtesy Bell System)

control circuits, will have been released, and the talking path will be maintained by the holding magnets of the crossbar switches, which are used on the link, district, office, and incoming link switch frames. These holding magnets are held operated under control of the supervisory relays in the district junctor and the incoming trunk circuits and are released only when the subscribers hang up their handsets.

The establishment of a connection with the crossbar switch is shown schematically in Fig. 29, in which 20 vertical units are connected to 20 subscriber lines and 10 trunks are strapped horizontally across the switch. With this arrangement, any one of the 20 lines may be connected to any one of the 10 trunks by closing the contacts at the proper cross point. By adding a second 200-point switch with 20 additional lines connected to its verticals, and extending the trunk strapping through both switches, 40 lines are given access to the 10 trunks. Thus, by adding other switches in this manner, the number of lines having access to these 10 trunks may be further increased.

A line link frame comprises primary bays and secondary bays. Each primary bay terminates 200 subscriber lines (10 primary switches with 20 lines each), but the number of primary bays per frame may be varied within limits to meet traffic requirements. The secondary bay contains secondary switches; the bay is divided vertically in the center, so that there are 10 switches, each with 10 verticals on the left of center and the same arrangement on the right of center of the bay. The switches on the left have their verticals connected to line junctors which are used for terminating traffic, and those on the right have their verticals connected to district junctors which are used for originating traffic. At the bottom of the secondary bay is a cabinet containing control circuit relays, and just above this cabinet are the multicontact relays which connect the control circuits to the crossbar switches.

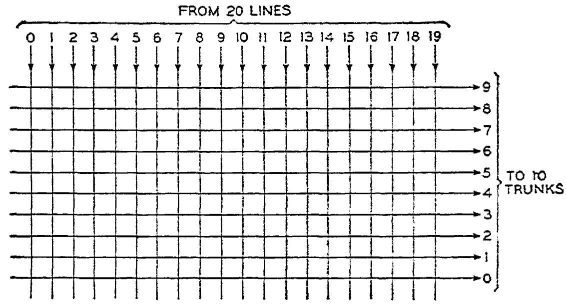


FIG. 29. Simple Trunking Arrangement with a Single 200-point Crossbar Switch—Crossbar System (Courtesy Bell System)

Since each subscriber in an office has only one crossbar appearance and that on a vertical unit of a primary crossbar switch, both originating and terminating calls are completed by means of the same line link circuits serving that particular switch. Thus, all originating traffic from any of the 20 lines on a primary switch flows through the associated 10 line links to the 100 district junctors, and all terminating traffic to these 20 lines flows through the same 10 line links from the line junctors, as shown in Fig. 30.

The arrangement shown in Fig. 30 is also used in the originating and terminating sender link switch frames, where the circuits reached are non-directional, that is, where any one of the circuits wired to the frame and available for selection can be used for establishing a connection.

Where it is necessary to provide greater flexibility and efficiency in trunk groups than is possible with the arrangement shown in Fig. 30, two primary and secondary switch

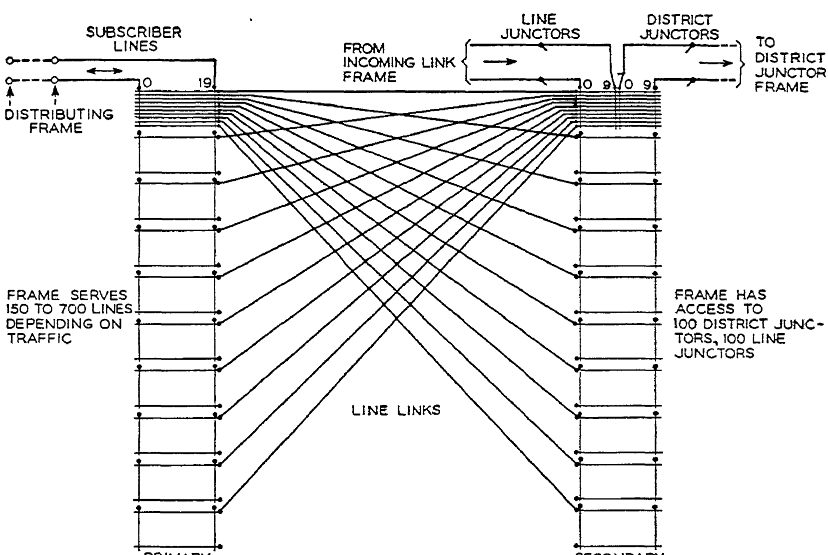


FIG. 30. Single Primary-secondary Trunking Arrangement—Crossbar System (Courtesy Bell System)

frames are employed, as shown in Fig. 31. This layout shows an incoming link frame to which incoming trunks are connected, and a line link frame to which subscriber lines are connected. These two frames are operated in tandem for establishing the terminating connections between the incoming trunks and are called subscriber lines. As shown, 100 incoming trunks are connected to the 100 horizontal paths of the 10 incoming link frame primary switches, 10 trunks per switch. Although only 200 lines (20 lines on each of 10 primary switches) are indicated in Fig. 31, 150 to 700 subscriber lines mav appear on the verticals of the primary switches of the line link frame.

Referring to Fig. 31, the connection of a particular incoming trunk to a particular called line requires the selection of an idle path through the incoming link and line link frames. This path will consist of an incoming link, a line junctor, and a line link. The incoming trunks on each of the primary switches have access to 20 incoming links appearing on the 20 verticals of the switch. These 20 incoming links are distributed over the 10 secondary switches of the incoming link frame, two links to a switch and one link to each half switch. In order to provide for the distribution of the 20 incoming links over the 10 secondary switches, the horizontal paths of the secondary switches are separated between the tenth and eleventh verticals, thus providing 20 instead of 10 horizontal paths on each switch. The incoming links on each half of these secondary switches have access to line junctors appearing on the verticals of these switches. These junctors are, in turn, distributed over

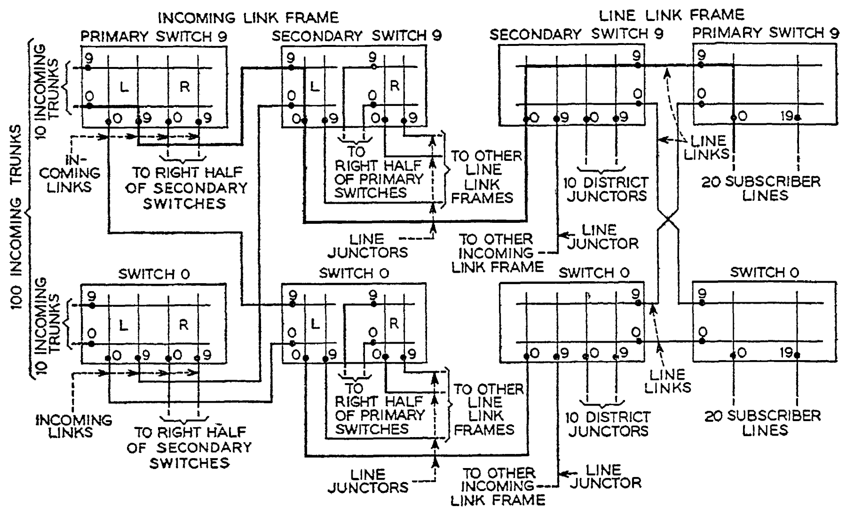


FIG. 31. Double Primary-secondary Trunking Arrangement—Crossbar System (Courtesy Bell System)

the secondary switches of all the line link frames in the office. There will be at least one junctor from each secondary switch on an incoming link frame to a secondary switch on every line link frame in the office, or a minimum of 10 line junctor paths between any incoming link frame and any line link frame, the number of paths varying, depending on the number of frames required in an office.

The line junctors on the verticals of each of the line link frame secondary switches have access to 10 line links on the horizontal paths. The 10 line links are distributed over the primary switches of the line link frame, one to each switch, giving each link access to the called subscriber lines appearing on the verticals of the primary switch with which the link is associated.

With the Fig. 31 arrangement of switches and the three groups of interconnecting link paths, any incoming trunk can be connected to any called line appearing on the line link frame or, by means of other groups of line junctors, to a called line on any other line link frame in the office. This trunking arrangement is also employed for connecting district junctors to outgoing trunks in the originating office.

Establishing a call from one crossbar subscriber to another crossbar subscriber requires four stages of operation, two at the originating and two at the terminating office:

1. The calling subscriber is connected to a subscriber sender through the line link frame, district junctor, and sender link frame, and the sender registers the dial pulses of the called number.

2. The subscriber sender is connected to an originating marker through the marker connector, and the marker then selects the switch frames for establishing a connection between the calling subscriber line and an outgoing trunk.

3. The outgoing trunk (incoming at the terminating office) is connected to a terminating sender through the terminating sender link frame, and the sender registers the pulses corresponding to the called number.

4. The terminating sender is connected to a terminating marker through a marker connector, and the marker then selects the switch frames for establishing the connection to the called subscriber line.

Future developments in crossbar equipment point toward simplification of equipment and circuits and reductions in the number of units required, wherever possible. An improved system of crossbar is now under study, in which it may be possible to establish connections with one marker instead of the two markers now used, and in which other simplifications may be secured.

OTHER RELAY SWITCHING SYSTEMS developed by various manufacturing companies include Relaymatic, by Kellogg Switchboard and Supply Co.; Relaydial, by Stromberg-Carlson Co.; All-Relay, by North Electric Manufacturing Co.

These systems operate on the principle of (1) finding the calling line when the calling subscriber takes his handset from the hook, (2) selecting and closing groups of contacts by relay action under control of dial pulses, subdividing the groups of contacts closed, until the contacts of the called line are reached, after which ringing power is applied to the called line by a link circuit. Since no moving parts are involved except the operating springs or reeds, which are equipped with precious-metal twin contacts, maintenance is simplified and operating costs are reduced over other types of automatic switching employing motors and step-by-step type switches with base-metal contacts and sliding brushes.

The line circuits may be assigned for common-battery local and rural, trunk or pay station service. Local lines may have individual stations or multiparty service up to 10-party selective (for metallic lines) or 16- to 20-party code ringing (for grounded lines). Local lines are of the metallic type, and line adapters are used for grounded rural lines, as required. By the addition of a trunk adapter, any line circuit may be converted to a trunk circuit. All link (connector) circuits have access to all lines, are assigned in rotation, and automatically release from any line in trouble or as soon as the subscribers hang up. A line lockout feature is generally provided with these systems which prevents tie-up of a link equipment if a line is in trouble or a handset is left off the hook or if the link fails to release after a predetermined time. As soon as the line in trouble is restored to normal, the lockout of the line from access to the link circuits is automatically discontinued.

These systems are made in capacities from 10 to 10,000 lines and may be operated as unattended small dial offices, if desired, with trunks to a nearby manual or mechanical office. In such cases, suitable alarms are provided which indicate at the attended office when the equipment needs attention and the type of trouble at the unattended office. Some systems of more than 200 lines employ relay-type selectors to distribute calls from one group of 100 line finders to the desired 100-line group of connector links and connectors.

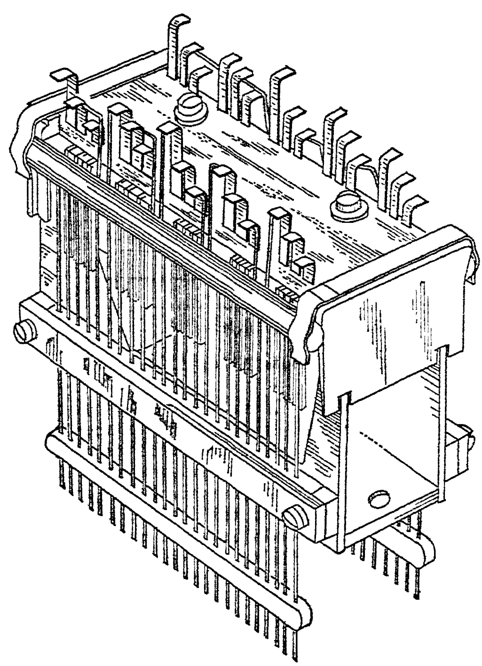


Fig. 32. McBerty Relay (Courtesy North Electric Mfg. Co.)

McBerty Automatic Telephone System (North Electric Manufacturing Co.). The important unit of this system is the McBerty relay, Fig. 32, a new design consisting of an integral reed spring-armature-contact structure having no pivots or hinges, which are common to most relays. The basic mounting structure may be used as a single multicontact relay or as a group of three, four, or five separate relays, the entire unit being light and compact. The relay coils are of the molded bobbin type, which slip over the steel-alloy cores welded to the mounting frame. Gold-alloy contacts are used throughout. Bare wire is used for multipling relay contacts.

This system employs a link circuit consisting of a line finder, connector control relays, and a connector for establishing a connection between a calling and called subscriber. As many links are provided as are required to handle normal traffic loads. Figure 33 shows a single link circuit layout for a 100-line system, but for purposes of illustration only three of the ten tens relays are shown. Also for clarity each single line shown represents two wires outside of and three wires within the switchboard.

When a calling subscriber removes his handset from the hook, the line relay in his line circuit operates, causing the proper line finder tens and units relays of an idle link circuit to operate, closing the calling line through to the connector control relays of this link. The calling subscriber then dials the called number, and the dial pulses are registered by

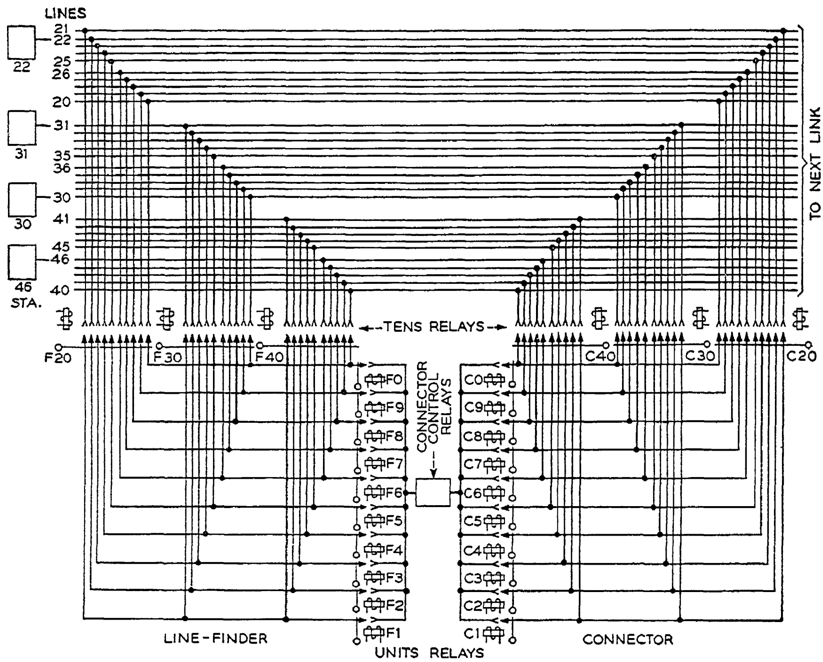


FIG. 33. Diagram of North 100-line All-relay System (Courtesy North Electric Mfg. Co.)

the control relays. These relays cause operation of the connector tens and units relays (corresponding to the called number) connecting the called line to this link.

The above operations complete the connection between the calling and called subscribers through the link circuit, which applies ringing power to the called line. This link is not released for other connections until the subscribers hang up. While the tens relays close through ten lines in both the line finder and connector each time they are operated, only one units relay is closed through in each line finder and connector for a given call, and thus all lines except the calling and called lines remain open to the control relays and may be seized by other links when calls are originated by such lines.

The XY dial telephone system, developed by Stromberg-Carlson Co., employs for its basic unit the XY selector-type switch, a view of which is shown in Fig. 34.

The XY switch, used in the line finder, selector, and connector circuits, is radically different in construction from the Strowger step-by-step switch. The switch is assembled on a metal base plate, which is mounted horizontally on a switch frame. The switch has a carriage with four separate wipers, tip, ring, sleeve, and hunt, and the carriage is moved as a unit in two horizontal directions, one paralleling the front edge of the base plate, called the X direction, and the other at right angles to the X direction, called the Y direction, under control of two magnets, X and Y respectively. The carriage is driven by a cog roller (tubular shaft) assembly, which slides along a shaft during the X motion and rotates during the Y motion. The cog roller is a double-cut tubular gear, with ratchet teeth cut parallel to its length and rack teeth cut as rings. The annular rack teeth mesh with and are driven in the X direction by a sprocket actuated by the X magnet. The Y magnet actuates a pawl which engages the ratchet teeth of the cog roller, turning it around the shaft.

Since the switch is 100 point, the X motion is given 10 steps (plus one for overtravel) and the Y motion 10 steps (plus one for overtravel), thus providing for selection of any one of 100 lines. Since four wipers are involved and each wiper has its own set of 11 wire banks (each 11 wires deep), 44 rows of wires are lined up in front of the wipers. An X wiper is also provided, which is operated by a pinion and rack assembly actuated by the X magnet and which has access to a 23-wire bank to mark the level of X travel.

When the X wiper finds the proper level, thus positioning the wipers, on the carriage, before the proper wire banks, the X magnet is de-energized and the Y magnet, assuming control, moves the wipers into the wire banks until the proper line wires are reached.

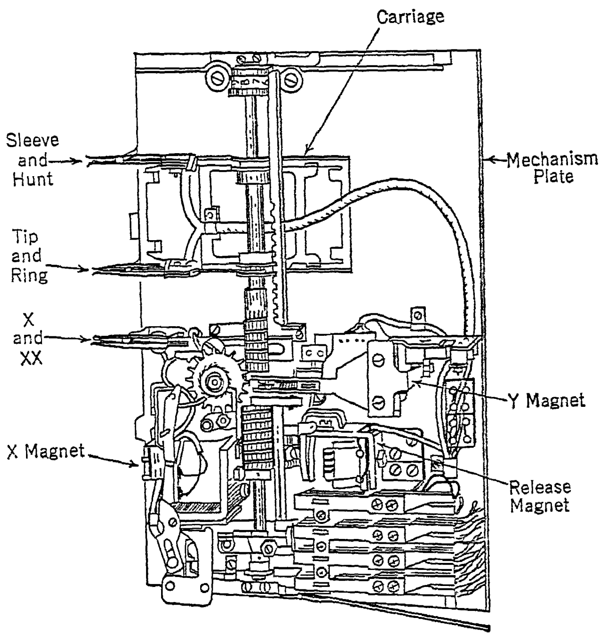


FIG. 34. The XY Switch (Courtesy Stromberg-Carlson Co.)

A number of unique features are built into the XY switch assembly, such as bare wire multiple for line terminations, a new mechanical design for magnet current interruptions to avoid armature chatter, and flexibility to function as a line finder, selector, or connector; the fact that the common wire banks need be wired only once, for up to 50 switches, results in large wiring economies.

The system requires line finders, selectors, and connectors, as in other step-by-step systems, but the XY switch functions for each of these three units.

In operation, the calling subscriber removes his handset from the hook, thereby causing his line relay to operate and an idle line finder, which is permanently associated with a selector, to find and connect to the calling line. The calling line is thus extended through to a selector which returns dial tone and supplies talking battery to the calling subscriber. Assuming that the called number is 234, the first digit dialed is 2, which causes the selector to move two steps in the X direction and into its wire bank automatically in the Y direction until an idle connector serving the 200 group of lines is found. When the second digit, 3, is received by the connector, it moves its wipers three steps in the X direction, and when the third digit, 4, is received, these wipers move into the wire banks in the Y direction to the fourth wire, thus connecting the calling to the called line. This connection is not completed, however, until the connector applies ringing power to the called line and the called subscriber has answered.

3. TOLL SYSTEMS AND OPERATION

Toll systems, as distinguished from local systems, are designed to handle toll traffic over toll circuits. Toll traffic differs very materially from local traffic, since toll circuits may extend from one toll office to a nearby toll office or to an office in this country or in almost any other country in the world, involving many thousands of miles of circuit.

Toll circuits must be of a grade suitable in all respects for the traffic they are required to handle. For the very long circuits, expensive equipment and complex arrangements are necessary to meet all requirements.

In the smaller offices the toll and local positions are identical or in the same lineup; the same line and cord circuits may be used for both local and toll service. In the larger toll centers, involving a number of toll circuit groups, separate manual toll boards are provided for concentrating in one switchboard, for a given toll area or center, all the toll circuits serving that area. Such switchboards require special circuits and auxiliary equipment for properly recording, ticketing, timing, and supervising toll connections.

MANUAL TOLL SWITCHBOARDS now in use in large metropolitan centers provide outward positions for outgoing toll calls, inward positions for incoming toll calls, and through positions for toll calls switched through the board from one toll circuit to another, or, if desired, toll positions may be designed to handle both inward and through toll calls.

One type of manual toll switchboard (Western Electric Co. No. 3C) now in use is equipped to handle all types of manual toll switching and, in connection with step-by-step offices, may be arranged to handle those local dial calls requiring the assistance of an operator.

The functions of this particular type of board are (1) to establish outward toll connections while holding the calling subscriber on the line (combined line and recording [CLR] traffic), or to establish connections later, if for any reason the call cannot be completed on the first attempt; (2) to connect inward toll calls to the called subscriber line if in the local or tributary office area; (3) to interconnect toll circuits (through traffic) upon request of a distant operator; and (4) to handle miscellaneous local calls.

Outgoing calls from local or tributary (small office, having toll connections to its larger toll center office) offices reach the toll board over recording-completing trunks from local manual or mechanical offices or over tributary toll circuits from tributary offices. Inward calls reach the toll operator over toll circuits (between toll centers) and are completed to the local called subscribers over toll switching trunks, either on a straightforward basis through manual B boards or by dialing or key pulsing over mechanical trunks through a mechanical office.

This switchboard has nine jack panels and three operator positions, each equipped with ten pairs of high-impedance toll cord circuits, per section, and by adding sufficient operator positions to handle the traffic load as many toll and trunk circuits as required can be terminated in the switchboard.

These boards are equipped with calculagraphs for stamping the tickets with the elapsed time of calls for billing purposes, electric clocks, ticket conveyors, ticket holders and compartments, and many other auxiliary devices for handling toll traffic. Transmission gain may be introduced in the toll circuits, as required, when the operator inserts the plug of her cord circuit into the toll line jack. Idle circuits and circuits busy may both be indicated by lamp signals, and automatic listening equipment may be provided at inward positions for incoming plug-ended toll circuit operation.

MECHANICAL TOLL SWITCHBOARDS OR SYSTEMS have been in use for some years in a number of comparatively small toll networks, the first systems being of the dial or step-by-step type.

The step-by-step system of toll dialing, though useful and economical for a small group of interconnected dial offices, presents sizable operating problems where intertoll dialing is attempted over an extensive area involving a large number of intermediate offices.

The mechanical switching of toll traffic requires that the digits dialed by a calling subscriber or originating operator at a given office, to reach a particular subscriber in another part of the country, must be different from the digits dialed to reach any other subscriber. In step-by-step toll dialing offices, the originating toll office is reached by the subscriber dialing zero (0), tributary offices by dialing the figure one (1), and two more digits are generally required to select the proper outgoing toll or tributary circuit. Thus, when a call is dialed through a number of step-by-step toll offices, three digits must be dialed for each office passed through in addition to the digits required for the terminating office and called subscriber line. This requirement results in a long series of dialed digits for a call that passes through a number of intermediate offices. Figure 35 shows that 19 digits are required for a call from Portsmouth, N. H., to Vinland, Kans., involving only four intermediate offices. Additional intermediate offices, in this connection, would increase the digits to be dialed, so that not only would there be delays in ascertaining the proper codes to dial in order to extend the call through the various intermediate offices, but also the dialing of a long series of numbers would retard operating time, would hold expensive facilities unnecessarily long, and would tend to increase operating errors. Also, if delays

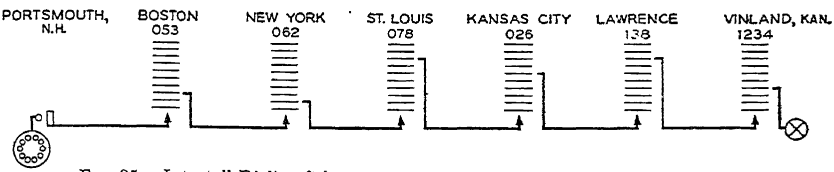


FIG. 35. Intertoll Dialing Scheme—Step by Step System (Courtesy Bell System)

were encountered in securing any intermediate link or if the called line was busy, the complete dialing process would have to be attempted again.

THE NO. 4 CROSSBAR TOLL SYSTEM, a development of the Bell System, was first placed in service in Philadelphia in August 1943. This system was developed primarily to provide for ultimate toll dialing on a nationwide basis and was designed so that it could be introduced gradually throughout the country on an economical basis without immediately displacing existing manual or mechanical systems except as desired.

The No. 4 system is arranged, as shown in Fig. 36, to complete (1) outward calls from local subscribers to outgoing toll lines, (2) inward calls from incoming dial or manual lines

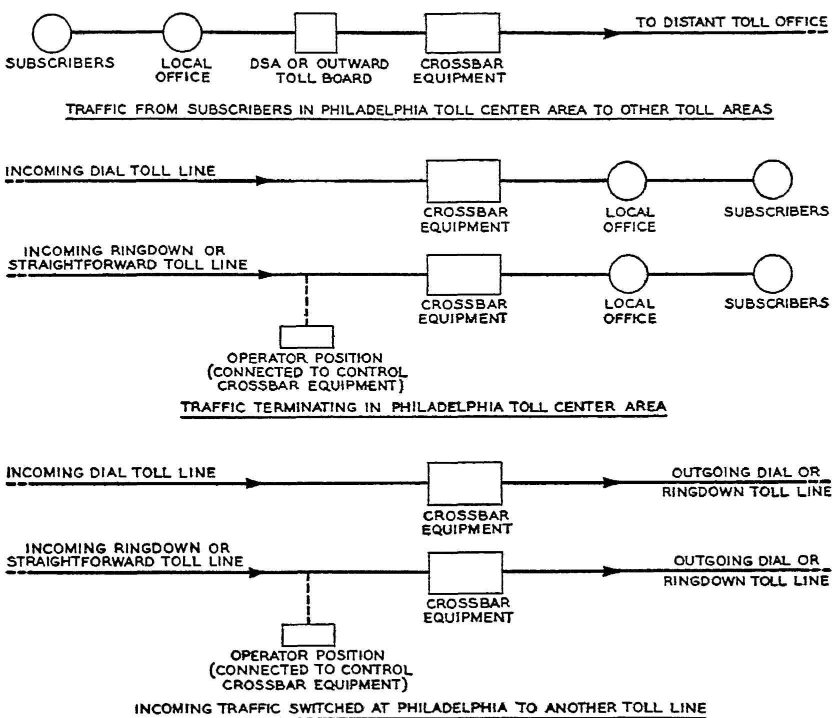


FIG. 36. Types of Calls Handled by the Crossbar System (Courtesy Bell System)

to local subscribers, (3) through calls between incoming and outgoing dial or manual toll lines.

Outward calls, and incoming calls from toll lines equipped for toll-line dialing, are automatically switched under control of dials or keysets at the originating end of the lines.

Incoming calls from other types of toll lines (ringdown or straightforward) are routed to operator positions.

In the No. 4 system, crossbar switches with senders and markers are used in the same general way as in the local crossbar system, with such variations as are required for toll traffic. The operator positions which supplement the mechanical switching system for handling terminating and through calls from toll lines not equipped for toll-line dialing are of the cordless type, but cord positions may also be used to facilitate handling calls over congested toll-line groups. This system permits toll-line dialing into a city having panel or crossbar offices.

Figure 37 shows a block schematic of the main circuit components of the No. 4 crossbar system. *Five types of senders* which act as automatic operators are provided, three for incoming and two for outgoing trunks. For each incoming call, an *incoming sender* is connected, and, unless the call is to be completed over a manual trunk or one equipped to receive multifrequency pulsing at the distant end (in which case only an incoming sender is required), an *outgoing sender* is also connected, as shown in Fig. 37. When an outgoing

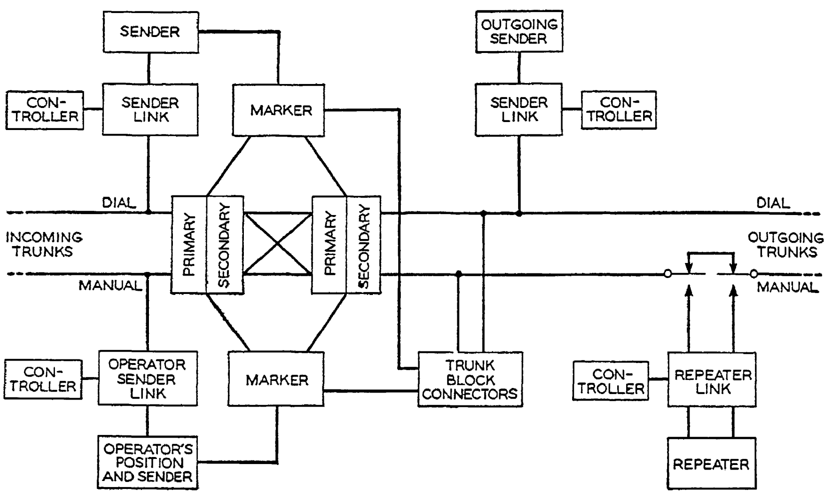


FIG. 37. Block Schematic of Main Circuit Components of the No. 4 Crossbar Toll System (Courtesy Bell System)

sender is used, the incoming sender transfers, through the primary and secondary frames, all the digits received except the first three, which are used by the marker to control the connection within the office. Incoming senders are designed to transmit d-c key pulses and outgoing senders to receive them, the transfer of digits being at the rate of 8 per sec. All incoming senders are also arranged to send out multifrequency pulses, so that outgoing senders are not required when the terminating points have senders capable of receiving this type of pulse, such as at local and toll crossbar offices. Ultimately, when multifrequency pulsing becomes general, outgoing senders may be eliminated. While incoming senders are arranged to send either d-c or multifrequency pulses, the *dial incoming sender* is designed to receive dial pulses (10 or 20 per second) recorded on crossbar switches, and the *key pulsing incoming sender* to receive d-c or multifrequency pulses, as may be indicated by a signal from the incoming circuit, as recorded on relays, four relays for each digit. The third type of *incoming sender* (*position*) is associated with each operator's position; on an outgoing call from the position, the sender connects to a marker, into which it passes the first three digits received, for the purpose of establishing connection to the desired outgoing trunk. When this trunk is selected an outgoing sender is attached, except when the trunk is on the manual or multifrequency basis.

Both types of *outgoing senders* receive d-c key pulses, but each type sends out different signals, depending on the signal the sender receives from the outgoing trunk. *One type* receives four or five digits and controls the sending of either revertive or call-indicator pulses. Revertive pulses are used for completing calls to panel or crossbar offices; call-indicator pulses, for calls to manual offices in panel areas. *The other type* of outgoing sender receives up to 11 digits, and either sends them out as dial pulses into a step-by-step office or connects itself to a call-announcer and controls the sending of the latter's voice announcements, which are limited to five digits.

Figure 38 shows the types of connections that incoming and outgoing senders are required to control. For a call to a local office within the crossbar toll office area, only the called office code and four or five digits are required to reach the subscriber from the crossbar toll office, since the trunk selected by the crossbar equipment connects directly with the called office. For a call to another switching area through an intermediate point, one, two, or three additional digits for use at the intermediate point must be dialed or keyed following the switching code, requiring up to 14 digits maximum.

All senders are safeguarded from being held too long on a connection by timing circuits, which, after a predetermined time, signal the originating operator to start the call again, and are then released. When trouble involving the sender exists, the sender is automatically held for inspection and the maintenance forces are notified by alarm circuits.

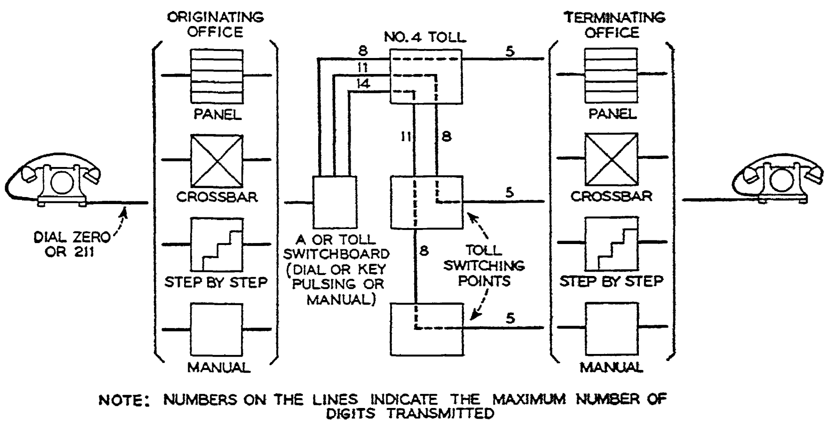


FIG. 38. Representative Types of Calls Switched by the No. 4 Crossbar Toll Office (Courtesy Bell System)

TELEPHONE REPEATERS are essentially voice-frequency amplifying devices with suitable talking and monitoring features designed for use in voice-frequency toll circuits which would, without amplification, be greatly restricted in their length. By means of such devices, properly spaced and controlled, toll circuits may be extended to any practical length desired.

In the early designs, telephone repeaters were basically of two types, those in the Bell System being designated 22-type for two-wire circuits and 44-type for four-wire circuits. Present Bell System practices employ a single type of amplifying device for either two- or four-wire circuits, and suitable connecting arrangements to connect it into the two- or four-wire circuits or any combination of such circuits.

Figure 39(*a*) and (*b*) shows a schematic diagram of the two- and four-wire arrangements, respectively. Where a repeater is employed at the junction of a two- and a four-wire circuit, the two-wire arrangement is used for connection of the repeater to the two-wire circuit and the four-wire arrangement for connection of the repeater to the four-wire circuit.

In the *two-wire arrangement*, the separate branches of the amplifiers are joined electrically through *repeating coil hybrids* for a repeater at an intermediate point in the telephone circuit. For a repeater at the terminal of a circuit a *resistance hybrid arrangement* (see Fig. 39[*c*]) usually takes the place of the repeating coil hybrid on the switchboard (drop) side of the repeater. In the *four-wire arrangement* repeating coil hybrids are not used, except where required at the junction of two- and four-wire circuits. Four-wire terminating sets are employed on the drop side of four-wire repeaters.

The repeating coil hybrid arrangement (Fig. 39[*a*]) consists of two repeating coils, *A* and *B*, with low-inductance windings so related and connected as to form, with associated equipment and the line, a balanced bridge circuit when the line and balancing equipment impedances are equal and the impedances connected to terminals 2–5 of each coil are equal. Coils of different ratios to match various line circuit impedances, and with phantom circuit taps for securing a phantom circuit, if desired, are available.

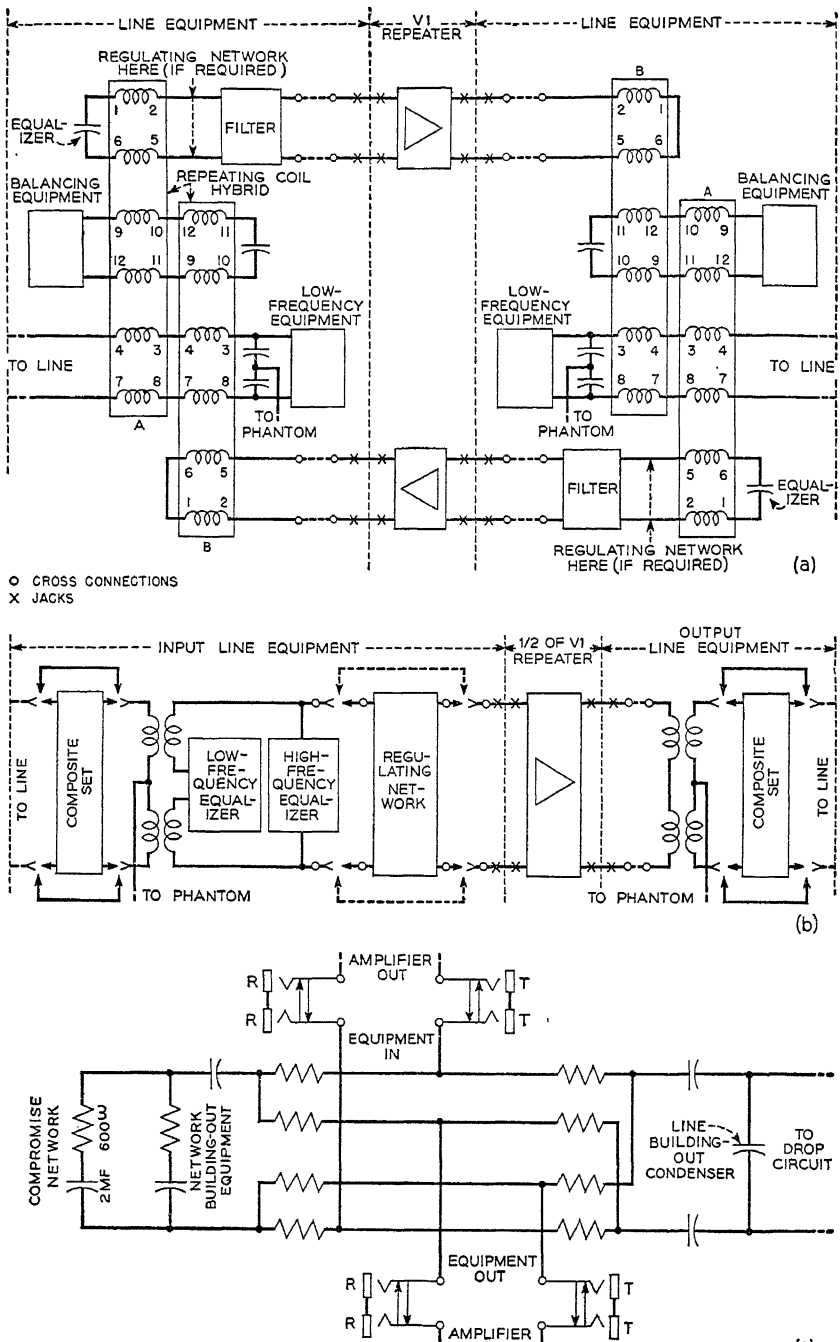


Fig. 39. Terminating Arrangements used with Telephone Repeaters (Courtesy Bell System)

Incoming current from the line flows through the line windings of both coils and induces equal voltages in the network windings 9–10–11–12 of both coils. Since the poling of the network windings of coil *A* are reversed with respect to the network windings of coil *B*, the resultant voltage across the network (balancing equipment) is zero. The voltages induced in the 2–1–6–5 windings of both coils are likewise equal, and the power received from the line (deducting coil losses) divides equally between the impedances of the amplifier branches connected to 2–5 of each coil. Since these branches transmit in opposite directions, the power received from the line from a given direction by one of the branches and its amplifier is effective, while that received by the other branch and its amplifier is ineffective.

Outgoing current from the effective amplifier flows through the 2–1–6–5 winding of coil *B* and induces equal voltages in the line and network windings of coil *B*. The resulting currents in these windings also flow through the corresponding windings of coil *A* to the line and balancing equipment, respectively. The currents in the line and network windings of coil *A* induce equal voltages in the 2–1–6–5 winding of coil *A*, but these voltages are opposing in phase because of the reversed poling of the network windings, and the resultant voltage is zero. The power received from the amplifier (deducting coil losses) thus divides equally between the line and the balancing equipment. The power received by the line is useful, while that absorbed by the balancing equipment is wasted.

In order to secure a satisfactory trans-hybrid balance, the windings of a given coil must be mutually balanced to a high degree of precision, but separate coils need not be so highly balanced with respect to each other.

The resistance hybrid arrangement (Fig. 39[*c*]) is composed of resistances, condensers, and the necessary terminating jacks connected to form a four-branch balanced lattice type network. This network joins together the amplifier branches on the terminating (drop) side of the repeater and terminates them in the required 600-ohm two-wire drop. The network presents a 600-ohm impedance in all four directions. The 1000-cycle loss through this network is 10.7 db for each direction of transmission.

In the *four-wire arrangement*, one-half of an amplifier unit is employed in each side of the four-wire circuit, transmitting in one direction only. Single repeating coils, of the type used with two-wire arrangements, are provided to match the impedances of the line and amplifier. Phantom taps are provided in these coils, and the coils must be matched for balance when phantom circuits are employed.

The amplifier circuit of the repeater is shown in Fig. 40. This circuit has a nominal input impedance of 600 ohms, and the input transformer impedance ratio (windings 9–8 to 1–7) is 300 to 357,000 ohms. The output transformer has four windings, of which winding 9–10 is for the negative feedback feature. The output impedance ratio of this transformer (windings 7–8 to 1–2) is 21,000 to 600 ohms.

The *amplifier vacuum tubes* are heater-type pentodes, 310 A for regulated and 328 A for non-regulated battery supply.

Grid bias on the tubes is obtained from the voltage drop in resistances *B* and *C* and in the *potentiometer* through which the total cathode current flows.

The total gain of each amplifier is about 35 db; the secondary winding of the input transformer is tapped to provide a total gain adjustment of 20 db in 4-db steps; and the potentiometer serves as an additional gain control with a range of about 5.4 db. The power-carrying capacity of the amplifier is such that the transmission level at the *amp out* jacks may be as high as +10 db with respect to the transmitting switchboard. The nominal d-c battery supply is 24 volts for the filament current and 130 volts for the plates.

Attenuation equalizers for two-wire repeaters are associated with the line equipments and are connected to the amplifier inputs at intermediate repeaters and to the receiving amplifier line input at terminal repeaters (terminals 1–6 and 2–5 of coils *A*, Fig. 39[*a*]). They are of the fixed type and designed for repeater sections of average length. The *low-frequency equalizer* for equalization in the low-frequency range consists of capacitance or a combination of capacitance and resistance, depending on the line characteristics. At two-wire circuit terminals, this equalizer is omitted in the transmitting side of the terminal repeater. *High-frequency* equalization on two-wire circuits is obtained by the effects of the various equalizer units mentioned and by interaction effects between the filter and the impedances between which it is inserted.

Equalization for four-wire circuits (see Fig. 39[*b*]) is provided by a low-frequency equalizer consisting of a condenser shunted by a resistance, and a high-frequency equalizer consisting of a combination of resistance, inductance, and capacity.

Low-pass filters of nominal 600-ohm impedance are provided, to limit currents above the voice range to be transmitted, in the four-wire branches of the line equipment, associated with the amplifier input for the two-wire arrangement, as shown in Fig. 39(*a*). Three types of filters are provided, having nominal circuit cutoffs of 2450, 2850, and 3500 cycles per second. The filter is omitted from the transmitting branch of terminal repeaters.

Regulating network equipment is provided for insertion in the repeater circuits (see Fig. 39[*a*] and [*b*]), as required, to compensate for changes in line attenuation due to temperature variations. This equipment functions under control of a pilot wire regulating system which actuates relays, causing resistance-type loss pads to be cut in or out of the repeater circuit, as required, to maintain circuit transmission levels. This equipment is more fully described in another part of this section.

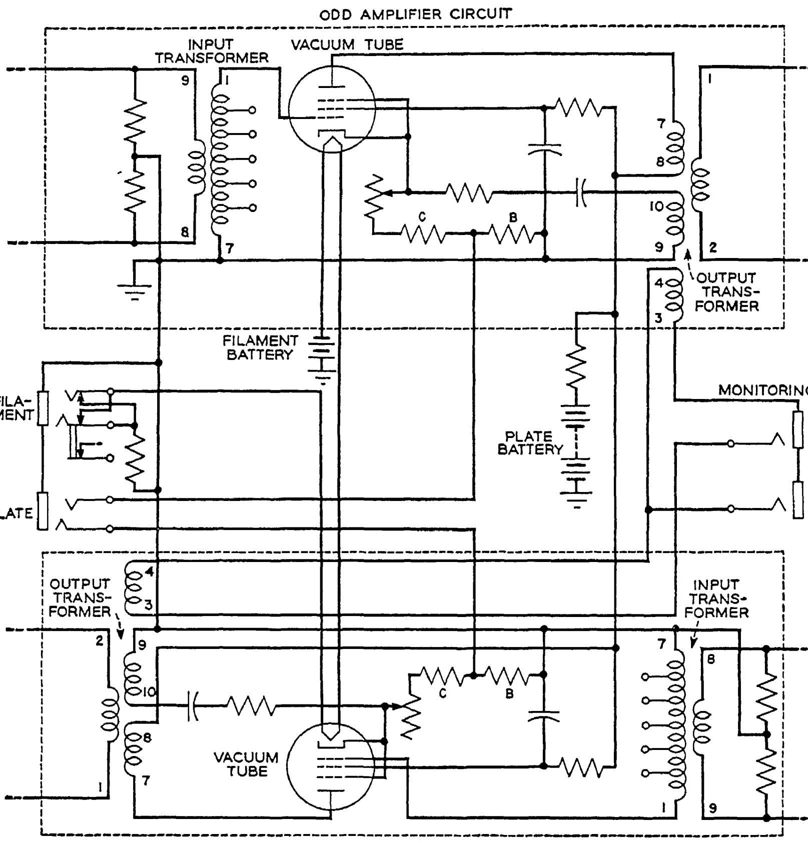


Fig. 40. VI Telephone Repeater Circuit (Courtesy Bell System)

The balancing equipment, connected to 9–12 of each *A* coil (Fig. 39[*a*]), is required to balance the line and its equipment (up to the repeating coil hybrid) in each direction of transmission for two-wire arrangements but is not required for four-wire arrangements, at intermediate points. At both the two- and four-wire terminal repeaters, a simple compromise network is employed on the drop side with the resistance hybrid (Fig. 39[c]) and four-wire terminating sets, respectively. Balancing equipments are designed in various combinations of resistance, capacitance, and inductance to closely match the impedance of their various associated lines.

Signaling over repeater-equipped circuits requires the use of auxiliary signaling circuits employing 20, 135, or 1000 cycles or composited d-c signaling on *two-wire circuits*. The 20- and 135-cycle and composited d-c signals are by-passed around intermediate repeaters, through which such signals will not pass, but 1000-cycle signaling will pass through the repeaters in the same manner as voice-frequency currents. For *four-wire circuits*, in which there are usually a number of repeaters, 1000-cycle signaling is generally employed as the most economical and satisfactory arrangement.

CARRIER TELEPHONE SYSTEMS permit the securing of additional telephone channels between two toll centers by superimposing carrier frequencies on voice-frequency wire circuits between these points. At the terminals of the carrier channels, carrier equipment is required which is capable of converting voice frequencies to modulated carrier currents, transmitting them over the wire circuit, and reconverting or demodulating them at the receiving end to voice frequencies. This equipment must operate in both directions of transmission. Carrier equipments (carrier repeater or transfer units) are also employed for amplifying carrier currents and for transferring the carrier channels where the wire circuit does not, but the carrier channels do, extend through the intermediate office.

Carrier telephone systems are in operation in many countries, but they are in use to the greatest extent in the United States because of its vast network of toll circuits. These systems are manufactured by a number of different companies, both Independent and Bell, the number of two-way channels provided in the various systems ranging from one to twelve, excluding the L-type carrier system. The Bell System is the largest manu-

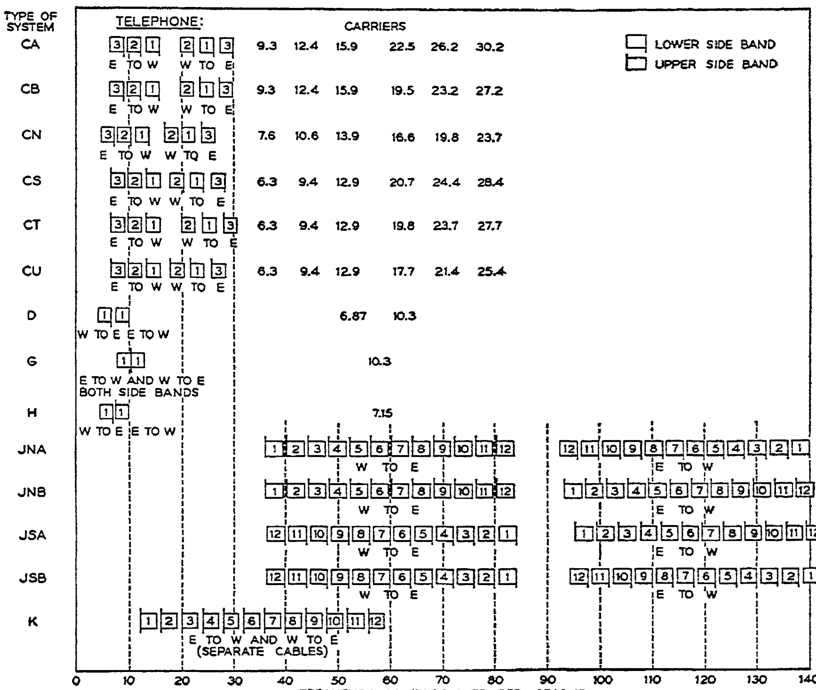


FIG. 41. Channel Frequency Allocations for Telephone Carrier Systems (Courtesy Bell System)

facturer and user of carrier telephone equipment in this country, and the frequency allocation chart shown in Fig. 41 applies to carrier systems of Western Electric Co. (Bell System) make. From this figure, it is seen that the D, G, and H systems provide a single channel, the C series three channels and the J and K systems twelve channels, all two-way.

In operation, the basic principle is the same, regardless of the type of system or the number of channels provided. Figure 42 shows a block schematic of a type C carrier telephone system with one intermediate carrier repeater.

It will be noted that each of the three channels has identical equipment units at both the east and west terminals, and that identical common equipment is provided at each terminal to serve each of the three channels. The carrier repeater also serves all three channels. Voice-frequency transmission over the wire circuit is not interfered with by, nor does it interfere with, the carrier currents, because of the low-pass and high-pass filters associated with the wire line at each terminal and at the repeater or transfer points. The low-pass filter will pass only voice and the high-pass filter only carrier frequencies.

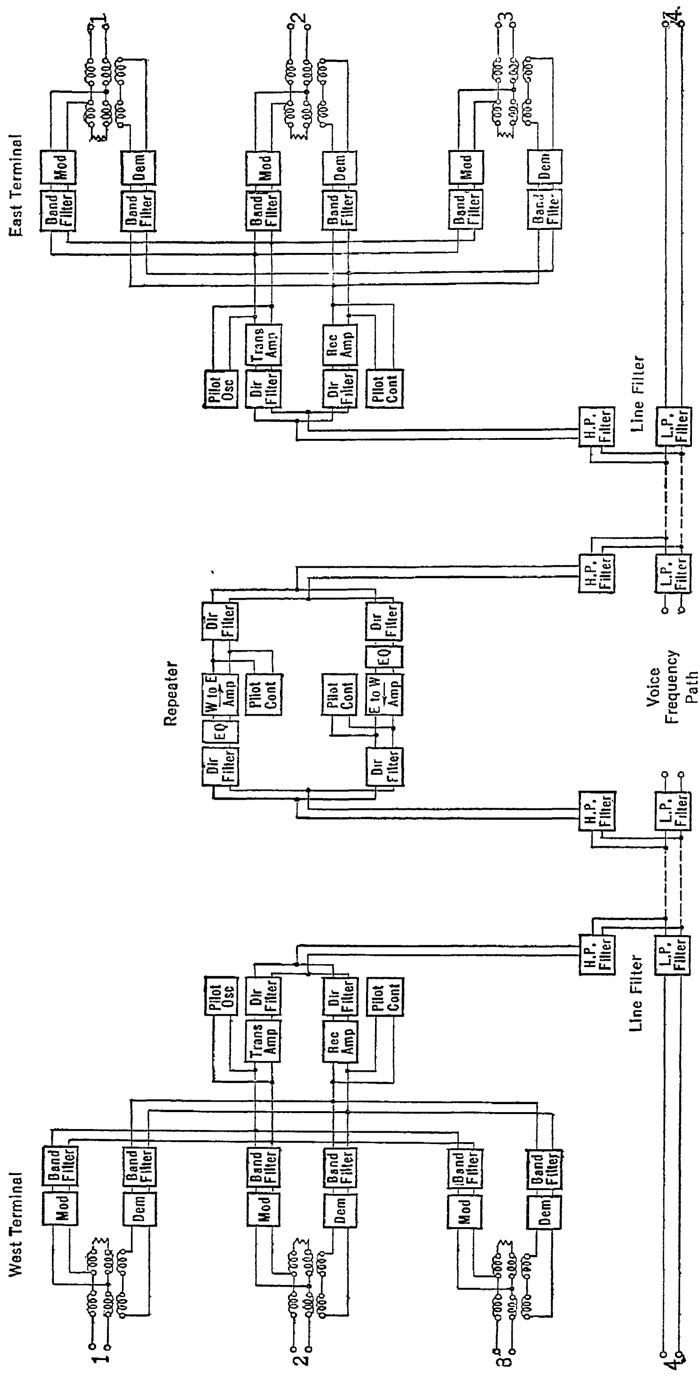


FIG. 42. Type C Carrier Telephone System (Courtesy Bell System)

Incoming voice frequencies (about 250–2750 cycles) into channel 1 (Fig. 42) at the west terminal pass to the modulator, where they modulate a carrier frequency. The lower sideband only (assuming this system is of the CS type) is transmitted through the band filter, transmitting amplifier, and directional and high-pass filters to the wire line. At the intermediate point the sideband frequencies pass through the high-pass and directional filters, equalizer, west to east amplifier, directional and high-pass filters, thence to the line. At the east terminal the sideband passes through the high-pass and directional filters, receiving amplifier, band filter, and demodulator to the voice terminal of channel 1. Transmission takes place similarly in the opposite direction. The band filters pass only the band of frequencies intended for their particular channels, blocking out all frequencies of other channels which travel the common paths. The directional filters separate the incoming and outgoing bands of frequencies. The amplifiers, which are of the high-gain, negative-feedback type, are necessary to maintain proper levels of transmission for the carrier currents. The pilot equipment indicates and controls the transmission levels automatically, so that manual adjustments are not required. The three-winding transformers (hybrid coils) separate the transmitting and receiving voice paths at the voice-frequency terminals of each carrier channel.

For *open-wire operation,* present practices make use of the G system for single-channel very short circuits (under about 25 miles), the H system for single-channel medium-length circuits (up to about 300 miles, with repeaters about every 125 miles), and the C system for three-channel groups ranging from about 100 miles to any length desired with repeaters about every 150 miles. The G system equipment does not have an amplifier, but one may be associated with the system externally. The CN allocation of the type C system and the D system are not used for new installations. All these systems are considered to be in the *low-frequency* group of carrier systems (up to about 30 kc), and they provide about 2500-cycle voice bands.

The *broad-band* carrier systems operate in a range from about 12 to 2000 kc or more and provide about 3000-cycle voice bands. As may be noted in Fig. 41, the K system, for toll cable use, functions between 12 and 60 kc; the J system, for *open wire,* between 36 and 140 kc; and the L system (not shown), for *coaxial cable* use only, between 60 and 2000 or more kc. The J and K systems may be operated any distance desired but require repeaters about every 70 and 16 miles respectively.

The L type carrier system, operating over specially designed conductors, known as coaxial cable, because of their construction, is capable of providing up to 480 two-way circuits per pair of coaxials (depending on the make-up of the cable) if coaxial amplifiers are spaced about 5 to 8 miles apart. Practically, however, the number of circuits to be provided in any L system will depend upon traffic requirements, since circuits can be added as desired at any time in groups of 12.

Figure 43 shows the frequency translations which take place at an L type carrier telephone system terminal. *Three steps of modulation* are employed to change an individual voice-frequency channel of 0 to 4 kc to its proper line frequency assignment. The *first step of modulation* occurs in the channel bank and translates a group of 12 voice channels to the 60 to 108 kc frequency band. The *second step of modulation* occurs in the group modulators and moves each fundamental group of 12 channels each (60 to 108 kc) to one of five frequency assignments (each 48 kc wide) within the 312 to 552 kc band. This band, designated a basic supergroup, is 240 kc wide and accommodates 60 channels of 4 kc each. The *third step of modulation* occurs in the supergroup modulators and translates each basic supergroup of 60 channels each (312 to 552 kc) to one of eight frequency assignments (each 240 kc wide) within the 68 to 2044 kc band. The 480 voice-frequency channels take line frequency positions within the 68 to 2044 kc band, and the four pilot channels are assigned to frequencies of 64, 556, 2064, and 3096 kc. The supergroups are separated by 8 kc each, except that 4 kc and 12 kc separate the first and second supergroups and the second and third supergroups, respectively.

All long-haul carrier systems, including the C, J, K and L systems, have regulating and pilot channel equipment which automatically adds transmission gain or loss as is necessary to maintain the channels within predetermined transmission equivalents. For *cable facilities* the normal transmission changes are due to temperature variations, being greater in aerial than in underground cable. Flat gain regulators with master controllers are employed in every repeater section for K systems. However, the amount of attenuation variation in cable pairs with a given change in temperature is not the same for all frequencies, an effect known as *twist.* To overcome twist effects, correcting circuits are also provided about every 100 miles for aerial and 200 miles for underground cable.

Large economies are possible with carrier systems, principally to provide additional toll circuits between toll centers. Their usage avoids, in many cases, the stringing of expensive wire circuits or possibly building a new pole line for open wire or cable or placing underground cable. In any event, studies will indicate the economies involved.

The basic 12-channel equipment units of the J, K, and L systems are similar, the channel modulators elevating the voice-frequency bands for 12 channels (4 kc per channel) from 0–48 kc up to 60–108 kc for these systems. Similarly the channel demodulator receives the 12 channels at 60 to 108 kc. The K system operates over separate toll cable pairs in separate cables with carrier repeaters, using the 12 to 60 kc band in each direction in the

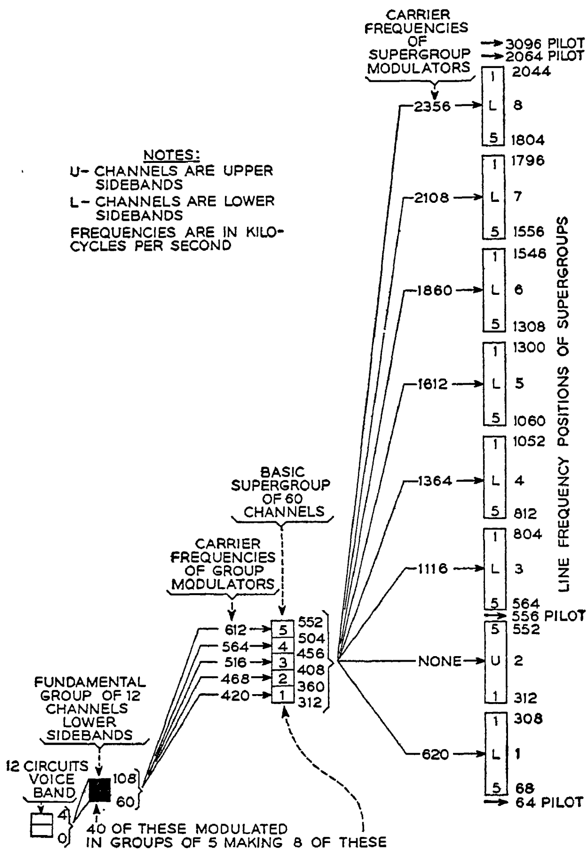


FIG. 43. Frequency Translations at an L-type Carrier Telephone System Terminal (Courtesy Bell System)

cables. The J system employs the same open-wire circuit in both directions, using line frequencies of 36 to 84 kc west to east and 92 to 140 kc east to west. The L system employs two coaxial units for the two directions of transmission, hence the same frequency band, from 60 to the required top kilocycle frequency, in each direction. Group frequency modulators move 12-channel groups from one frequency range to another, as required.

Figure 44 shows a block schematic of a type J carrier telephone terminal; Fig. 45, a type K carrier telephone terminal in more detail; and Fig. 46 a frequency translation diagram for the type J system.

The particular range of frequencies, 60 to 108 kc, to which the voice frequencies are first elevated in the J, K, and L systems was chosen, because (1) high-grade crystal filters can be most economically built for operation in this general range, (2) the second harmonic (120 kc) of the lowest frequency (60 kc) lies well above the highest frequency (108 kc), precluding the possibility of interference between the second harmonic (from any channel) and other channels, and (3) manufacturing economies are achieved by using the same

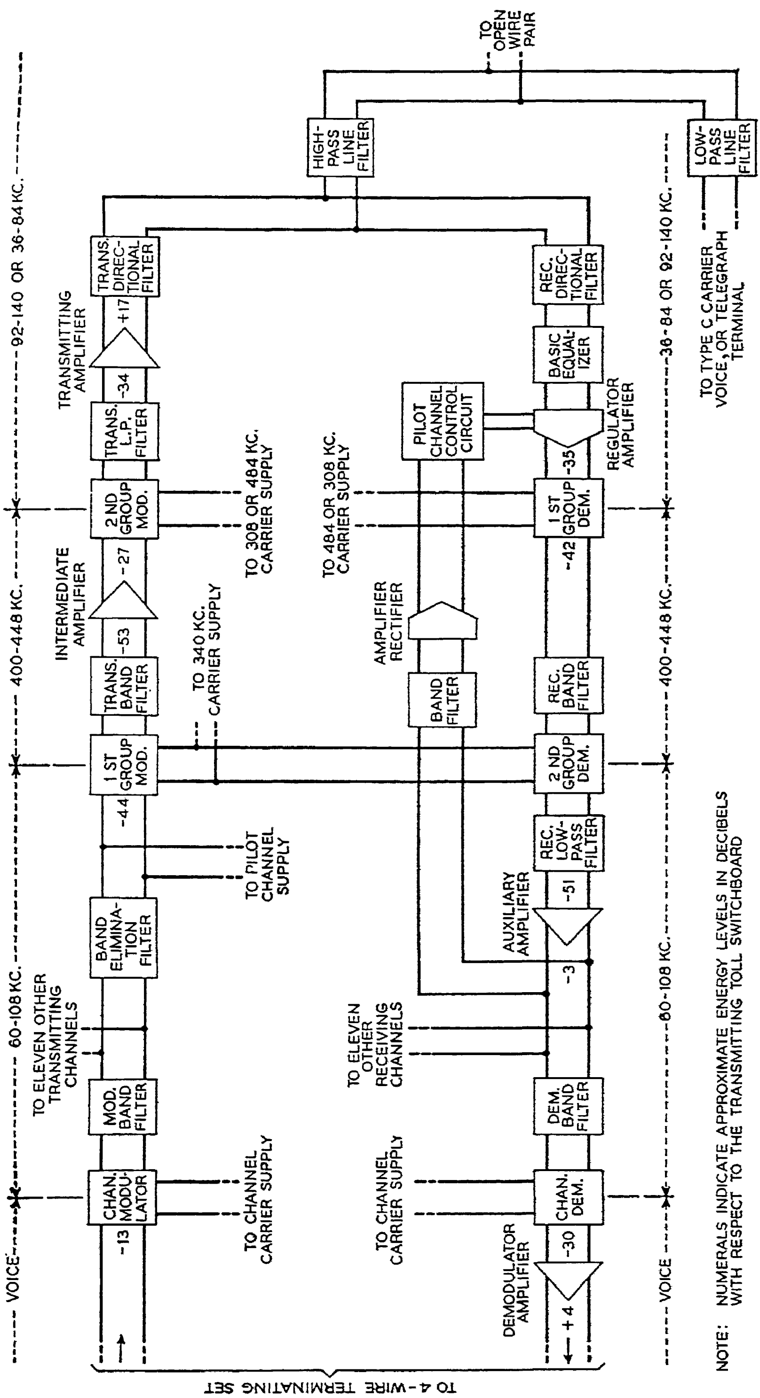


Fig. 44. Type J Carrier Telephone System Terminal (Courtesy Bell System)

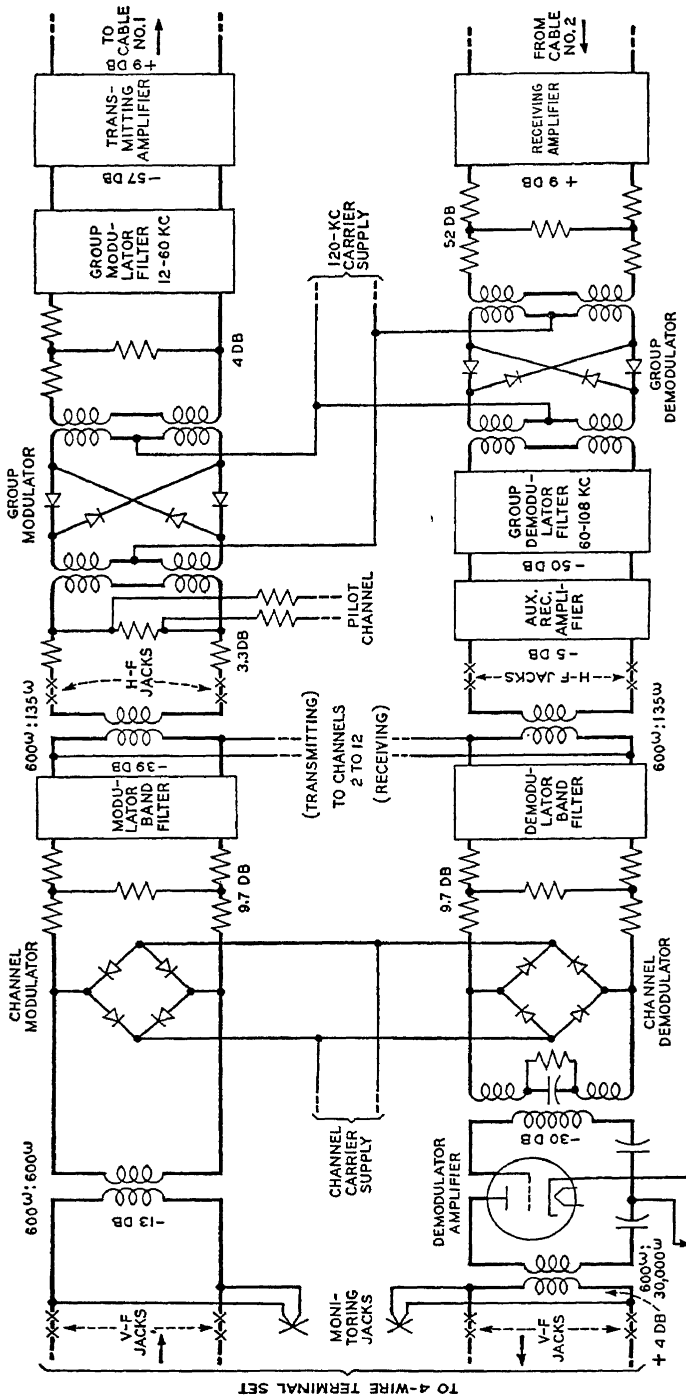


FIG. 45. Type K Carrier Telephone System Terminal (Courtesy Bell System)

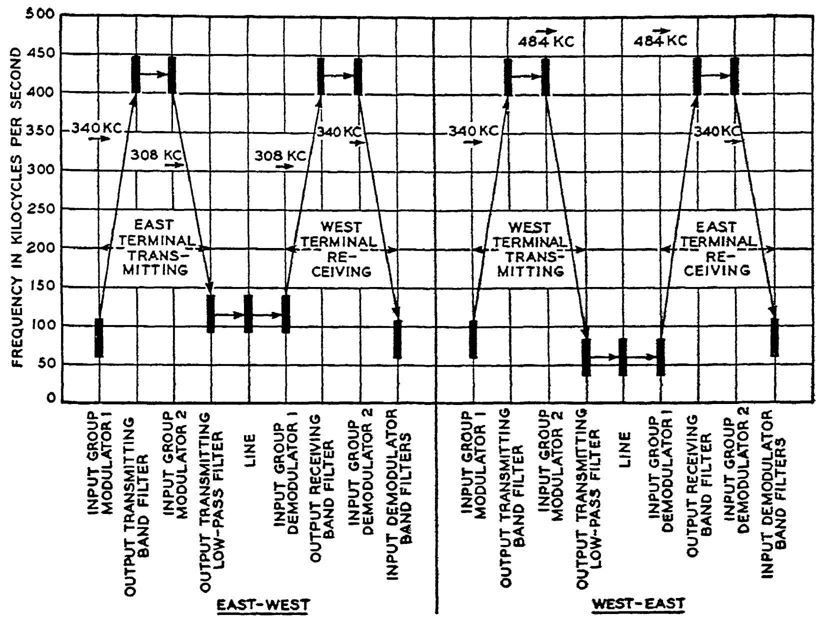


FIG. 46. Frequency Translations in Type J Carrier Systems (Courtesy Bell System)

group of channel carrier frequencies for all the broad-band systems. The line frequencies 12 to 60 kc were chosen for the K system because, for non-loaded cable pairs (loading not being available for these frequencies), the attenuation increases with frequency and in the frequency range chosen the attenuation-frequency increase is about uniform.

4. TANDEM SYSTEMS AND OPERATION

Local tandem offices are provided, usually one to each extensive multioffice area, where a large number of offices are located within and adjacent to city boundaries. The tandem office is generally centrally located to the offices it serves and has direct trunks to all these offices and also to attended PBX (private branch exchange) switchboards having a number of pay stations, to the toll board, and to various special service boards. It is neither economical nor practical in large metropolitan centers to provide a group of outgoing and a group of incoming direct trunks between each pair of offices in the center, since cable plant and the terminating trunk equipments required would be too costly. For many combinations of two offices some distance apart the traffic is usually so light as to preclude the use of expensive cable pairs between them for trunk purposes. Thus, the tandem office provides an economical means of trunking calls in the larger centers because a common trunk group is provided between each office served and tandem.

Although the introduction of a tandem office in the call trunking system of a given area adds another office to the system and may materially lengthen the trunk mileage between certain pairs of offices, depending on the geographical location of the tandem office with respect to the surrounding offices, it is evident that two groups of trunks (one outgoing and one incoming) between each pair of offices in a large center of, say, 50 offices would result in an uneconomical trunk network and equipment layout and a more complex arrangement of handling traffic. Also, the greater number of calls handled by the tandem trunk groups increases their efficiency and may largely offset the tandem layout costs. Tandem offices are warranted, however, in any particular case only if a comprehensive study of all the various factors involved indicates their need.

One system of tandem trunking now in use provides for trunking from a local manual office to a *panel sender tandem office* on a straightforward basis. The local A operator, upon receiving a call, inserts the calling plug of an A cord circuit into an idle outgoing trunk to tandem, causing selection to be made at the tandem office of an idle operator and a tone to be sent back to the A operator indicating that the tandem operator is ready to receive the call. The A operator passes the called office name and number over the trunk to the tandem operator, who registers this information on a keyset (a unit composed of individual keys having office code and number designations), with which each tandem position is equipped. The operation of the keys in the keyset causes pulses or number announcements to be sent out over the tandem to called office trunk, thus reaching the called office.

If the called office is of the step-by-step, panel, or crossbar type, the incoming pulses from tandem reach selector (in step-by-step) or sender equipment (panel or crossbar) in the called office, which completes the connection to the called subscriber in the regular manner, as previously described for these systems.

If the called office is of the manual type, the incoming pulses reach call announcer equipment at the tandem office before passing out over the trunk where they are converted from pulses to spoken numbers which reach an idle B operator's ear at the called office; or, where call announcer equipment is not provided, the tandem sender equipment sends out pulses over the trunk to call indicator equipment at the called office, which causes the called number to be displayed before the B operator. In either case the B operator connects the plug-ended trunk circuit over which the call is being routed into the called subscriber multiple jack, and ringing automatically starts on the called line. Supervision of the call at the local, tandem, and B boards is by means of the usual lamp signals.

If the calling subscriber is in a mechanical office (step-by-step, panel, or crossbar), the call is usually routed to a special board in the calling office. The special operator passes the call to tandem on a straightforward basis, and the connection is then completed by the tandem operator in the same manner as described for a calling local manual office.

Crossbar tandem equipment is now standard for new tandem offices, rather than the manual tandem arrangement, just described, where it is applicable. This equipment handles calls over two-wire trunks from panel or crossbar offices to other panel, crossbar, or panel indicator manual offices by means of crossbar switches in a marker system of operation. Calls from a manual office through crossbar tandem would reach a tandem operator over straightforward trunks and be completed as described for the panel sender tandem office.

The major switching frames in crossbar tandem offices correspond somewhat to the frames in a local crossbar office. Incoming trunks, terminated on incoming trunk frames, connect through trunk link frames, an office junctor grouping frame, and office link frames to outgoing trunks, which terminate in incoming trunk equipment in other offices as shown in Fig. 47.

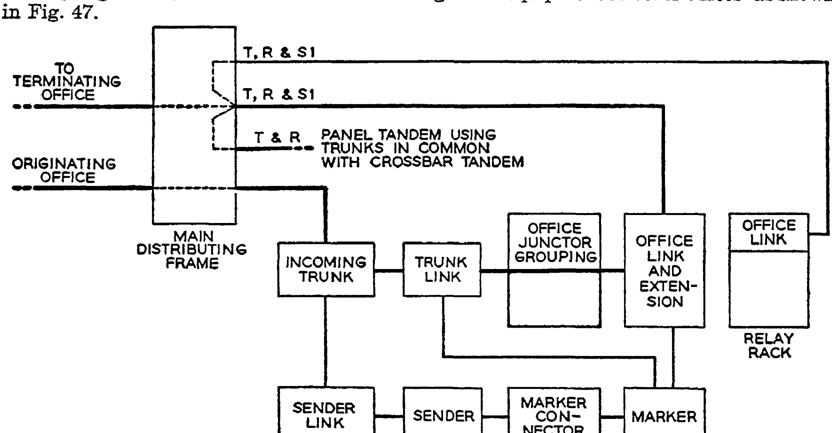


FIG. 47. Crossbar Tandem Office Units—Crossbar System (Courtesy Bell System)

When a call reaches the tandem office incoming trunk equipment, tandem sender and marker units are caused to associate with the trunk and link frames involved in the call. The sender receives and registers the incoming pulses from the originating sender on a revertive (pulses sent back) or dial pulse basis, and it controls the tandem switch selections and the selections in the called office. The marker which is associated with the sender through the marker connector frame routes the call through the tandem equipment under control of the tandem sender registrations.

5. AUXILIARY SERVICE EQUIPMENT

SERVICE OBSERVING DESKS are provided, where required, at individual offices or at a central location for observing the performance of switchboard operators and switching equipment.

One type of service observing desk is intended primarily as a non-centralized observing bureau for use with toll or local crossbar, step-by-step and manual, or combinations of toll and local plant. This desk is generally employed where not more than one desk position is required, and observations are confined to circuits at the same location as the desk. This desk is equipped with a cord circuit, operator's telephone circuit, jacks, lamps, keys, and other equipment for observing on lines and trunks.

For mechanically operated systems, pen registering equipment is also employed to record the subscriber dial and line registering pulses, as required.

Central observing desks require direct lines to the various offices which are to be observed, and as many desk positions as necessary to make the desired observations.

The information obtained by means of these desks is a very important aid in improving operator and equipment efficiency and in bettering service generally to the public.

INFORMATION DESKS are provided at individual offices or central locations to furnish subscribers with information about telephone numbers not listed in the telephone directory or changed from the directory listing and about many other items essential in assisting the subscribers to secure a wholly satisfactory telephone service. Up-to-date files of telephone listings are maintained at the information desks accessible to each information operator. Centralized information desks require that subscribers in each office be routed over the trunking system, provided for such service, to the centralized desk. Individual office information desks are reached over direct intraoffice trunks.

CHIEF OPERATOR DESKS are, if required, located one in each office. Various types of calls are referred to the chief operator, who is in direct charge of the operating forces. Complaints from the subscriber about the service rendered or any other items regarding the wishes of the subscriber in connection with calls the handling of which is not the operator's function are referred to the chief operator or a supervisor.

Trunks are provided from manual or DSA boards to the chief operator desks, over which these calls are routed.

REPAIR SERVICE DESKS are required in all but the very small offices to receive subscriber complaints of service or reports of trouble encountered with the substation equipment or telephone plant in general. Many other types of reports from the public are also referred to these desks, which are convenient contact points for subscribers. Records are maintained of troubles reported and cleared on subscriber lines, and valuable data are secured from them for studies of troubles and their elimination.

Repair trunks are provided at individual offices or to a central point for receiving these calls except in small offices where such calls are usually received at local testboards.

6. COMMON SYSTEMS

MAIN DISTRIBUTING FRAMES (MDF) are required in central offices for terminating the outside local and toll lines, which are usually brought into the office on cable pairs. These pairs terminate directly on protector strips, mounted vertically on the vertical side (VMDF) of the frame. In the smaller offices local and toll MDF are usually combined; in the larger installations, involving a number of toll and toll entrance cables, separate local and toll frames are provided. Terminal strips with insulated metal terminals are mounted horizontally in rows (shelves) on the horizontal side (HMDF) of the frame, and these terminals are cabled to an intermediate distributing frame (IDF). Cross-connecting wire (jumpers) may be run between any vertical protector strip and any horizontal terminal strip so that any incoming line on the VMDF may be connected to any pair of terminals on the HMDF. Other arrangements of protector and terminal strips are also used, depending on the needs of the individual office.

INTERMEDIATE DISTRIBUTING FRAMES (IDF) are usually employed in both manual and mechanical offices for local and toll lines. These frames have terminal strips on both the horizontal (HIDF) and the vertical sides (VIDF) of the frame. For *manual local lines* cabling from the MDF is terminated on the HIDF terminal strips, to which cabling from the manual A and B boards is also multipled. For small subscriber lamp multiple and single office boards, the B board and its cabling are omitted. The subscriber answering jacks in the A board and the line circuit equipment are cabled to the vertical terminal strips (VIDF). Cross-connecting wire may be run between any horizontal and any vertical terminal strip, thus providing a means of connecting any outside line to any subscriber A and B board multiple jack and to any answering jack and line circuit. By properly distributing heavy and light calling lines throughout the switchboard, traffic loads can be more uniformly spread over the operator positions.

Relay racks are also provided for mounting various types of apparatus, such as relays, repeating coils, apparatus mounted on panels, testing equipment, and many other equipment units.

PROTECTORS are provided at the vertical side of the MDF in all central offices having exposed outside cable or open wire plant in order to protect the equipment from damage due to excessive voltages and currents from foreign sources such as lightning and power lines. Figure 48 shows a typical protector used at main frames. Where the outside cable enters aerially and is exposed to these foreign sources, as it may be in small offices, fuses are also required in the circuits, unless the entering cable has at least 6 ft of 24 or finer gage cable inserted in it in such manner that no power line contacts can occur between the point of insertion and the MDF.

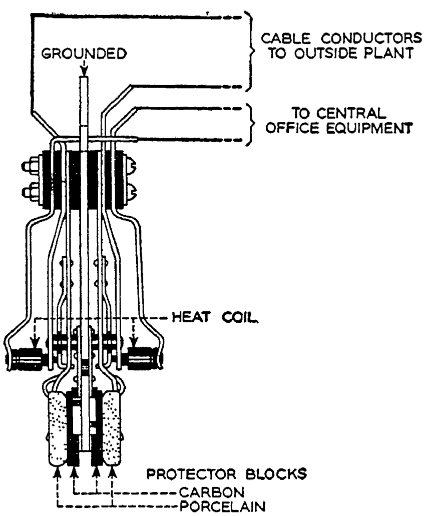


FIG. 48. Main Frame Protector with Heat Coils and Protector Blocks (Courtesy Bell System)

The protector consists principally of a spring assembly, arranged to hold two heat coils and two sets of protector blocks, one of each for each side of a metallic line. The heat coils designed to protect delicate central office equipment usually operate if 0.35 amp flows for 3 hours or longer or if 0.54 amp flows for more than 210 sec. This coil consists of a small coil of wire wound around a copper tube, into which is inserted a metal pin held in place by easily melting solder. The coil of wire is placed in series with the line conductor, and if sufficiently heated the solder melts, releasing the metal pin, which is forced against a grounding spring by the outside line spring of the protector. Several types of protectors and heat coils are available; some of the heat coils open the line conductor when they operate and may be reset mechanically without replacing or resoldering the coil. The protector blocks consist of a porcelain block with carbon insert and a solid carbon block, a pair to each side of the line, the air gap between the insert and carbon block being about 0.003 in. This gap will, on the average, break down at about 350 volts potential.

The protector ground bars are bonded to a main-frame ground bar which is connected to the office ground.

TESTBOARDS are provided in all attended central offices, and testing equipment of suitable design in all offices, for the purpose of determining the condition of lines and equipment, either because of reported or indicated trouble or as a trouble-preventive measure.

Many types of test cabinets and testboards have been made available to meet the needs of the particular location in which they may be installed.

In *small manual local offices* test cabinets, equipped with voltmeter (with batteries), test keys and cords, trunks, and in some cases an external Wheatstone bridge, are employed. In *large manual installations* a number of positions of testboard may be used, depending on the volume of repair and subscriber installation work involved. Each of these positions usually has a voltmeter, test keys and cords, trunks, and access to one Wheatstone bridge, as required. Lines and equipment are tested on reports of trouble or after installation work is completed, and, in case of trouble, its location is determined and the trouble is cleared by outside repairmen.

In *manual toll offices* testboards (primary, secondary, or both) are arranged with a group of jacks for each toll circuit, which is wired through these jacks before reaching the toll switchboard. The test positions are provided with voltmeter, Wheatstone bridge, test keys and cords, and trunks for testing lines and equipment and locating and clearing troubles. The testboard jacks also serve to patch lines and equipment to spare lines and equipment when the regular layout is in trouble or to make temporary circuit changes.

In *step-by-step local or toll offices* the testboard equipment is comparable to that provided in manual offices except that the positions are equipped with dials and dialing trunks for connecting to the various lines and equipments.

In *panel dial and crossbar local offices* a local test desk test selector frame is employed, consisting of a bay for sequence switches and a bay for relays, with terminal strips located at the side of the frame. This frame is used to associate the local test desk with the final selectors which are arranged for testing subscriber lines. The operation of the cutoff relay in the subscriber line circuit may be controlled from a key in the test desk. The six first selector circuits which connect to trunk circuits in the test desk have access to any one of 15 second selectors, which terminate in final test selectors. Test trunk ringing circuits are furnished as required. The test desk is equipped with voltmeter, test keys and cords, trunks, Wheatstone bridge if desired, and other necessary apparatus for checking the condition of and locating and clearing trouble on lines and equipment. In addition to the test desk there are various frames for testing the operating condition of the equipment, such as the outgoing trunk test frame, decoder test frame, and multiple registration test unit.

In *crossbar toll offices* a toll test board is provided for making overall tests on the toll trunks in order to locate troubles and restore the trunks to normal. A jack field is provided through which the intertoll trunks are wired, and miscellaneous trunks are terminated in this field. The circuits in this testboard are four-wire, requiring twin plugs. For talking and monitoring, the circuits are reduced to two-wire by means of a hybrid coil. A transmission-measuring system with the readings projected on screens at the ends of the testboard, a noise-measuring unit, and a variable oscillator are provided as part of the testing equipment. Outgoing toll trunks may be locked out (made busy) and tested by dialing through the *trouble tracing frame*, which seizes the incoming trunk connected to the outgoing trunk in trouble or to be tested and lights a lamp associated with that trunk at the testboard. The desired outgoing trunk may also be seized by plugging a test cord into the test jack appearance of the incoming intertoll trunk and operating the lockout relay in the outgoing trunk through the connecting switches.

In addition to the toll testboard a *maintenance center* is provided for each No. 4 crossbar toll system, in which various testing frames are located with a chief switchman's desk and files in front of the frames.

The maintenance forces at this center are occupied with responding to alarms, making tests, following up trouble reports and assistance requests, and maintaining records of the operations.

7. POWER SYSTEMS

Power equipment for central offices is required to provide direct current for talking and signaling and alternating and pulsating current for signaling and for many other auxiliary needs in telephone operations.

Power equipment for *magneto offices* usually consists of dry cells or a battery eliminator to supply direct current to the operator's telephone set and a hand generator and power-operated ringing device for ringing subscriber bells. Magneto subscriber telephones are supplied transmitter current by dry cells at each telephone or other power sources.

Power equipment consists of motor-generator sets and rectifier units (in *common-battery offices*) for supplying direct current for the energization of subscriber telephone transmitters, private branch exchanges or private automatic exchanges, central office cord and operators' circuits, relays, switches, alarms, carrier systems, telephone repeaters, and many other central-office units. Continuity of service is insured by the provision of storage batteries which float across the d-c office power supply and, in case of commercial power failure, will carry the office load for a short period of time, and by duplicate charging units operated by other than commercial electric power.

The types of charging and load supplying power units (including storage batteries) and of the signal supply units vary over a wide range of equipments, which have different capacities, functions, and characteristics depending on the office power load demands and the purposes for which the units are designed.

All common-battery offices, both local and toll, require 24- and 48-volt d-c supply, which may be provided by regulated rectifiers (mercury vapor, tungar, copper oxide, or selenium types) for the smaller loads; and motor-generator sets, singly or in multiple units, for the larger loads. In all cases, storage batteries of suitable capacity to provide for the 24- and 48-volt demand are bridged across the d-c power supply leads on a float basis; that is, the office load is carried by the generating units, the batteries serving as an emergency source of power. These offices may also require a 130-volt plate supply for carrier systems, telephone repeaters, and other vacuum-tube devices, which supply may be furnished from regulated rectifiers for small loads or from motor-generator sets for the larger loads. Storage batteries of the proper capacity are also required across this supply on a float basis for emergency reasons. Figure 49 shows a typical central-office power-plant arrangement with a single floating battery and automatic control. Although only one motor-generator set is shown, it is the usual practice to add these units for multiple operation as required.

Ringing and signal supply units may be of the vibrator or subcycle converter type for small offices or motor-generator sets for the larger offices. These units are designed to supply the usual 20-cycle, 75 to 175 volt ringing power, but certain multifrequency sets

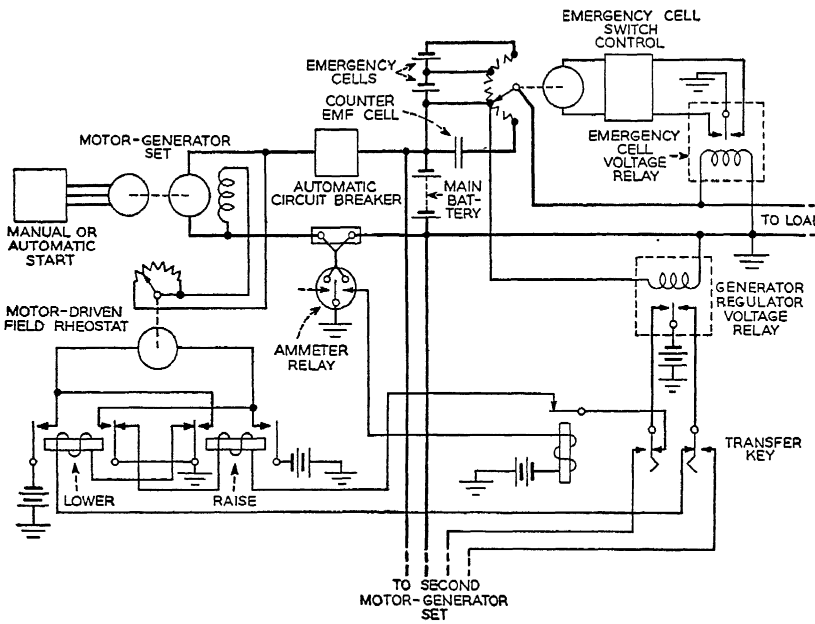


FIG. 49. Typical Central Office Power Plant Arrangement with Single Floating Battery and Automatic Control (Courtesy Bell System)

are also available for supplying harmonic frequencies of 16 $^2/_3$, 33 $^1/_3$, 50, 66 $^2/_3$, and, if desired, 25 cycles, at from 75 to 175 volts, for selective party-line ringing. These units also provide 110 to 120 volt d-c coin control supply, superimposed ringing with 46 to 52 volt silent interval tripping battery supply, howler tone (applied to lines with receivers left off the hook), interrupter tones of various frequencies for operator and subscriber signals, such as busy, trouble, or operating procedure.

In addition to the above types of signaling units there are several different types of single and multifrequency generators producing frequencies for signaling, such as 135 and 1000 cycles, and for vacuum-tube oscillators.

All power equipment is automatically regulated to close limits.

RADIO TELEPHONE SYSTEMS

Radio telephone systems have been in operation for over 25 years, but with the growing need for such systems in recent years, accelerated by World War II, they are rapidly coming into use for many purposes in the communications field. Many of these systems are arranged for connection to wire telephone plant, so that subscriber stations may be interconnected over radio channels and wire line extensions to give local, national, and world-wide service.

8. APPLICATIONS

Radio telephone service is being employed today (1) between all of the larger and many small countries, (2) within individual countries, (3) between ships at sea and fixed land stations, (4) between coastal harbor and inland waterways ships and fixed land stations, (5) between mobile vehicles of various classes and stationary points or other mobile units, (6) between planes and ground stations or other flying planes, (7) directly between persons (walkie-talkie), and (8) for special and emergency use, such as for fire, police, emergency repair units, and for temporarily bridging gaps in telephone lines that have sustained major damage.

Development work and trial tests are in progress for the use of radio channels (1) between and within trains and railroad operating control points, and (2) for rural line telephone service. An important activity, now in the experimental stage, is the use of super high-frequency (microwave) radio systems for toll telephone, television, facsimile, and other services. Undoubtedly many other applications of value to the public will be found for radio telephone systems.

Radio telephone systems may be listed at present under the following general classifications:

1. Long-haul toll.
2. Short-haul toll.
3. Coastal harbor and inland waterways.
4. Highway mobile.
5. Urban mobile.
6. Airways mobile.
7. Railway mobile.
8. Rural subscriber.
9. Special emergency.

Table 1 lists some of the principal operating data regarding these systems.

A new *microwave radio relay system*, using frequencies between about 2000 to 12,000 or more megacycles, is installed between New York and Boston. This system, using "line-sight" frequencies, requires seven intermediate relay stations, spaced on the average about 30 miles apart and located on relatively high elevations, as shown in Fig. 1. This trial of

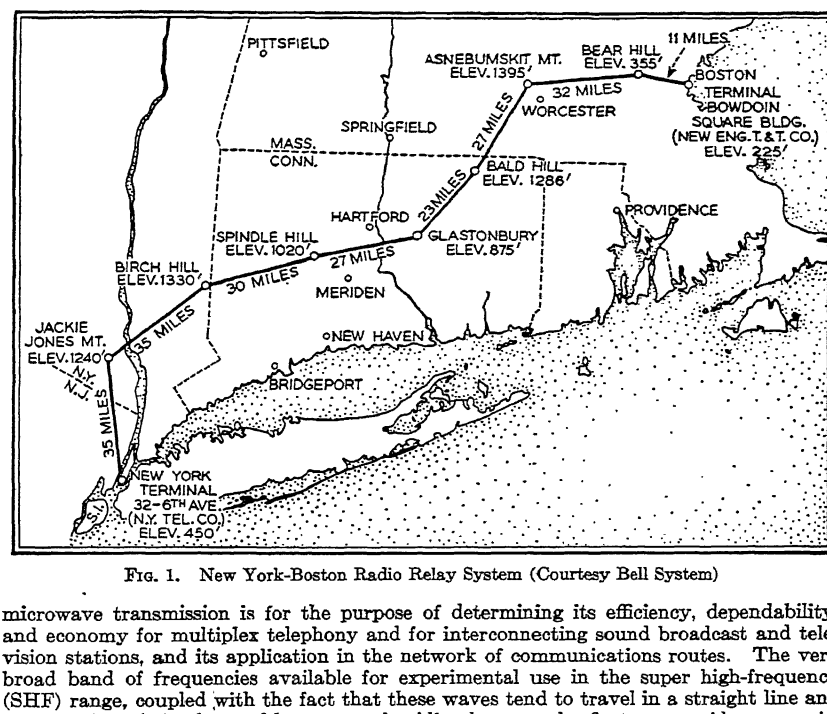


FIG. 1. New York-Boston Radio Relay System (Courtesy Bell System)

microwave transmission is for the purpose of determining its efficiency, dependability, and economy for multiplex telephony and for interconnecting sound broadcast and television stations, and its application in the network of communications routes. The very broad band of frequencies available for experimental use in the super high-frequency (SHF) range, coupled with the fact that these waves tend to travel in a straight line and are capable of being beamed by means of guiding lenses and reflectors, provides a promising field for experiments in communications, with the probability of eventually securing

Table 1. Radio Telephone Systems—Approximate Ratings

Type of Service	Frequency Range,* megacycles		Reliable Operating Range, miles	Transmission and Operational Methods Used				Transmitter Rating. Carrier or Sideband Maximum, kilowatts	Transmitting Radiation System. Net Gain Referred to Vertical Half-wave Doublet, decibels	Receiving Antenna Noise Improvement Factor, decibels	Receiving System Sensitivity Provided, decibels above 1μv/m	Minimum Field Intensity, decibels above 1 μv/m	
				Two Sidebands and Carrier	One Sideband, Suppressed Carrier	Constant Volume †	Constant Net Loss					30 Mc ‡	150 Mc
Long haul toll	Long wave......	0.05– 0.07	2000–3000		x	x		50 –150		25 to 50	−5 to +5		
	Short wave ‖....	4 – 25	1000–9000	x		x		2 – 80	+6 to +18	6 to 18	−10 to +10		
	To ship.........	4 – 21	0–4000	x		x		2 – 80	0 to +12		+5 to +20		
	From ship ¶.....	4 – 21	0–4000	x		x		0.25 – 2.0		0 to 12	−5 to +20		
Short haul toll ‖		2 – 25	100–1000	x		x		0.25 – 5.0	−3 to +15	0 to 18	−5 to +20		
		30 –3000	10– 200	x			x	0.01 – 0.1	+10 to +20	10 to 20	+20 to +50		
		152 – 162	10– 50	x			x	0.01 – 0.05	9	9	+10 to +40		
Coastal harbor (over sea water)	To ship.........	2.1 – 2.8	300	x		x		0.4			+20 to +30		
	From ship ¶.....	2.1 – 2.8	300	x		x or	x	0.05			+10 to +20		
Inland waterways (over land or fresh water)	To ship	2.1 – 2.8	50	x		x		0.4			+20 to +30		
	From ship ¶.....	2.1 – 2.8	20	x		x or	x	0.01 – 0.05			+10 to +20		
VHF marine	To ship.........	30 – 44	50	x		x		0.25				−5 to +30	
		152 – 162											+10 to +35
	From ship ¶.....	30 – 44	50	x		x or	x	0.02				−5 to +20	
		152 – 162											+10 to +25

Highway mobile	To vehicle......	30 - 44	0–20	x		x		0.05 - 0.25				−5 to +30	
	From vehicle....			x		x or	x	0.015- 0.03				−5 to +20	
Urban mobile	To vehicle......			x		x		0.25					+10 to +35
	From vehicle....	152 -162	0–20	x		x or	x	0.015					+10 to +25
Airways mobile	To plane........			x		x							+10 to +35
	From plane.....	108 -132	**	x		x or	x						+10 to +25
Railway mobile	To train........			x		x		0.015- 0.25					+10 to +35
	From train......	152 - 162	0–20	x		x or	x	0.015- 0.025					+10 to +25
Rural subscriber	To subscriber....			x			x	0.25	0 to 9	9			+10 to +35
	From subscriber.	152 - 162	0–25	x			x	0.015	9	0 to 9			+10 to +25
Special emergency		1.6 - 3.2 25 - 28 30 - 44 152 - 162	0–30 0–20	x		x or	x	0.015- 0.5					

* FCC Allocation Table for 25 to 30,000 Mc, revised to July 19, 1946, and FCC Proposed Allocation Table for 25 Mc and below, revised to July 12, 1946.

† Method applies to land station.

‡ These values assume half-wave dipole or whips. The lower value assumes set noise controlling; the higher value corresponds to "high" noise conditions which are not expected to be exceeded more than a small percentage of the time.

x indicates method used.

‖ Microwave (1750 to 30,000 Mc) in experimental stage.

¶ Quiescent operation from ship to shore.

** Extremely variable, depending on altitude.

large groups of telephone, telegraph, television, sound broadcast, and other useful channels.

The multicavity magnetron, a high-frequency power generator developed for radar, has made possible 10,000-Mc frequency currents with peak powers ranging from 10 to 1000 kw for very short intervals of time.

The microwave radio relay system has been given a valuable tool in the *lens-antenna*, consisting of an array of small metal plates mounted in a frame about 10-ft square. This lens employs the same principle in focusing radio waves into a pencil beam as an optical

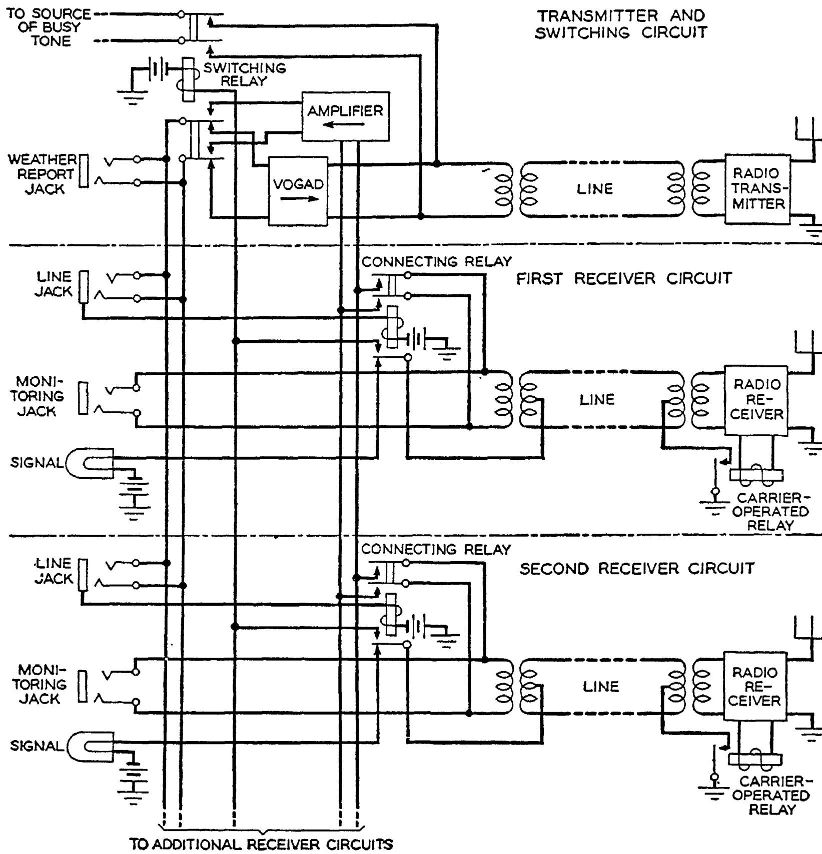


Fig. 2. Shore Circuit for Coastal Radio Service (Courtesy Bell System)

lens does in directing light waves. The radio waves are fed into this new lens-antenna through a hollow-tube wave guide and horn at the rear, the waves spreading out along the hornlike shields to the lens plates, which bend the waves and direct them out of the front of the lens to produce an emergent wave front parallel to the front of the lens. This type of antenna causes less wave distortion due to dimensional variations than would result from such variations in a parabolic reflector, and greater manufacturing tolerances therefore are permissible. Also, the horn-type shield at the rear of the lens reduces rearward radiation, present in reflector antennas, and the front of the lens may be protected from weather by a plastic covering.

Multiplex operation over microwave radio channels was developed by the Bell Laboratories during World War II for use of the U. S. Army Signal Corps and proved to be of great value to the armed forces. Its full possibilities for commercial service are being explored. Army-type pulse modulated microwave systems are now in commercial service between Los Angeles and Catalina Island as well as from the mainland to Nantucket Island. As developed for war purposes, multiplex operation employed a highly directive and sharply focused microwave beam of about 5000 Mc which carried eight separate messages. The intelligence of each channel is conveyed by varying the time position of the 1-microsecond channel pulses, eight of which (one for each channel) are transmitted in sequence, 8000 times a second. Thus, if a 1000-cycle tone is being beamed over one channel, one cycle of this tone requires $^1/_{1000}$ of a second, during which time eight pulses, spaced at approximately equal intervals throughout the one cycle, would be transmitted. At the receiving terminal these pulses are received in sequence by their respective channel and, being representative of the electrical intelligence at the originating terminal, are reconverted into sound intelligence. Two-way operation is obtained by using a separate radio channel in each direction of transmission. Because of the method of transmitting, this system has been designated as *pulse-position modulation* (*PPM*).

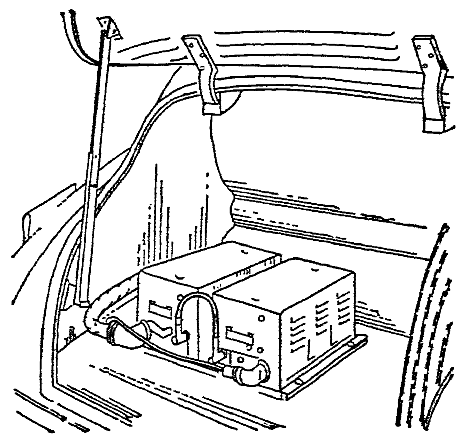


FIG. 3. Mobile Radio Transmitter and Receiver in Auto Trunk (Courtesy Bell System)

Coastal harbor and inland waterways radio telephone systems are extensive, one consisting of 14 basic shore stations, strategically located to cover the entire coast from Maine to Florida, to Texas, and up the Pacific coast to Seattle. Figure 2 shows a circuit of one of the *shore* stations. Note that by use of the monitor jacks the shore operator can select the receiver giving the best reception.

Highway mobile radio telephone systems give the operator the same facility. These systems consist principally of (1) f-m (frequency modulation) radio transmitters (about 50 to 100 miles apart) and associated receivers at fixed locations along the intercity highway which the system is to cover, (2) f-m radio transmitters and receivers for the mobile units, and (3) a control terminal associated with each fixed transmitter.

Wire lines connect the fixed receivers to the control terminal, which provides for linking the transmitter and as many as eight receivers to a two-way, two-wire line to the central office handling the system.

In operation, a customer desiring connection with one of his mobile units moving along a highway between his city and a distant point which is covered by a radio system asks for the mobile service operator in his city. This operator has access over a wire circuit to the various radio transmitters along the highway and, on the basis of the probable location of the desired mobile unit as furnished by the customer, a code assigned to the particular vehicle is dialed and is transmitted by wire and by the selected radio transmitter to the called vehicle. This code activates a selector set in the vehicle, which gives a visual and audible alarm. In calling from the vehicle, the occupant removes his handset from its holder and presses his push-to-talk handset button, which causes his transmitter to send out a signal. This signal is received by the nearest fixed receiver along the route being traveled by the vehicle, which converts the radio- to a voice-frequency signal. This signal then passes along the receiving wire circuit to the terminal equipment at or near the telephone office and signals the mobile service operator.

FIG. 4. Control Unit and Hand Set of a Mobile Radio Unit Mounted under Auto Instrument Panel (Courtesy Bell System)

Figure 3 shows a mobile installation of a transmitter and receiver in the trunk compartment of an automobile; Fig. 4, a control unit and handset under the dashboard of an automobile; and Fig. 5, a schematic of a two-tone selective signaling mobile unit.

Urban mobile radio telephone systems employ equipments like those in the highway mobile service for transmitting and receiving telephone messages or signals. However, since such systems are intended to cover only an individual city and its adjacent territory (up to about 20 to 25 miles from the transmitter), usually only one transmitting station is provided at a central location in the city and a number of receiving stations are strategically placed within the area to be served, so that the mobile transmitter signals will be received and carried to the central office in the city, wherever the vehicle may be in the area. One additional feature not provided in highway service is a one-way signaling service: when certain signals are received by the occupant of the vehicle, certain action is to be taken. Voice communication is not given in signaling service.

The fixed f-m transmitter will generally have a power output of 250 watts using an assigned frequency in the 152–162 Mc band. The mobile f-m transmitter will usually have a 15-watt output at a different frequency in this band. Initially, mobile units will be limited to one channel for a given city, and a number of such units will be assigned to the same channel.

Selective signaling units, actuated by coded dial pulses transmitted from the fixed transmitter and installed in the vehicles, will insure signaling only the vehicle being called. A two-letter, five-digit code will be used, such as WU 2-5556, which will not be duplicated

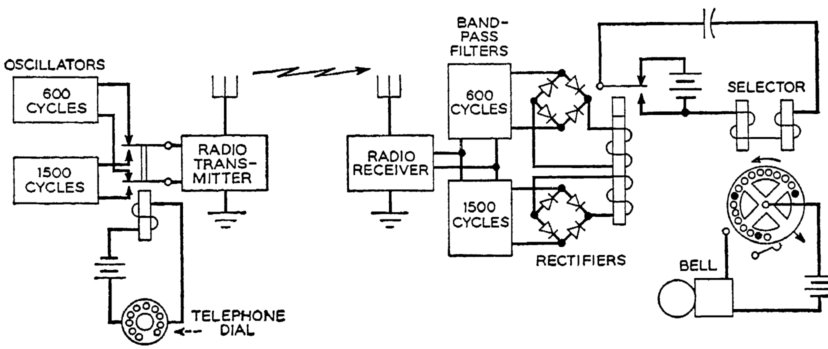


FIG. 5. Two-tone Selective Signaling System for Mobile Radio Service (Courtesy Bell System)

on the same frequencies, where interference might result. Since the mobile selector set operates the calling signal on the twenty-third pulse, the sum of the digits (not including letters) will always add up to 23, unless the plan of selection is later changed. Differentiation between vehicles operating only in one local area and those that may operate in more than one such area will be effected by assigning different code letters to the two classes of vehicles.

Railway mobile radio telephone systems are now (1949) under trial tests by several railroad companies (1) between trains and railway control points, (2) within the train itself, and (3) for use of the passengers in talking to fixed telephones.

The last-mentioned service makes use of highway mobile installations, which in most cases parallel the right-of-way. Service between passengers on moving trains and land telephone systems may be rendered by carrier-induction methods, rather than by radio, or by both types of transmission, as may seem best in the future.

Rural subscriber radio telephone systems are now in the trial stage, employing assigned frequencies within the 152 to 162 Mc band and low-power transmitters.

At Cheyenne Wells, Colo., a system has been installed (operating on a trial basis in the frequency range of 44 to 50 Mc) which provides for rural radio service to eight ranches located from 11 to 21 miles from the town. None of these ranches are reached by either power or wire telephone lines, the operating power for the ranch radio sets being obtained from home electric plants. Four of the ranches are served by direct radio links to the Cheyenne Wells central office; the other four are reached by relatively short wire extensions from a nearby radio-equipped ranch. At the central office the eight stations are grouped, through terminal equipment, to form an eight-party rural line.

Radio equipment at each of the first four ranches includes a transmitter, receiver, a telephone set, and two antennas. The 10-watt transmitter and associated receiver are housed in a steel cabinet, which can be located out of sight, with only the telephone instrument in view. The antennas are mounted on a pole or atop one of the ranch buildings.

At Cheyenne Wells, the equipment includes a 50-watt transmitter, receiver, and terminal equipment for associating the various radio links with the central-office switchboard.

Further studies will be made, based on the results of the trial tests, to determine the most practicable types of radio equipment to employ for rural service.

Special emergency radio telephone systems have been in operation for many years to serve fire, police, forestry, highway, utility companies, and a number of other services. Operation is of low power output at frequencies generally within the 1.6–3.3 and the 30–40 Mc bands.

The equipment employed for the mobile services is, in general, similar to the types discussed under coastal harbor, highway, and urban mobile service with variations required for the particular type of service involved.

The telephone and other wire-using companies have been using for several years portable emergency radio telephone equipment to bridge gaps in open-wire lines resulting from sleet, flood, or wind damage.

One type of portable radio telephone system for emergency use employs a 50-watt transmitter arranged to operate at one of ten selected frequencies within the 2.0–3.1 Mc band, and a receiver adjustable for any frequency within the 2.0–3.1 Mc band. One antenna at each terminal is switched between the transmitter and receiver, depending on the direction of transmission, by a voice-operated control unit. A volume limiter in the transmitter, and automatic volume control and a codan (carrier-operated device, antinoise) in the receiver, regulate the transmitter modulation and the receiver output and operation to provide uniform transmission. Its operating range varies with the type of terrain, noise, and atmospheric conditions from about 25 to 50 miles. Power for the terminal units is obtained from 110–120 volt, 60-cycle supply. A sensitivity control circuit actuates the transmitter at a minimum level of −47 dbm (47 db below 1 milliwatt) and likewise prevents line noise or room noise from causing false operation of the equipment.

9. TRANSMISSION AND OPERATIONAL METHODS

Quiescent transmitter operation is employed either to save power or to permit the installation of a single transmitter and receiver at the same location under conditions which otherwise would prevent satisfactory communication. The transmitter is provided with manual or voice-operated control means which render it sufficiently inactive during idle intervals so that emission from it does not interfere with reception. If saving power is an objective, the switching is directed toward securing minimum power input during quiescent periods.

Two sidebands and carrier transmission is the most commonly used method, owing to the simplicity of signal generation and detection. It requires a radio transmission band equal to twice the highest audio frequency to be transmitted. The carrier contains no intelligence-bearing signal component but simplifies detection and is useful for such control purposes as automatic tuning and volume control at the receiving end and for the operation of auxiliary relays. The power required to transmit the carrier is large compared with that required for the sidebands, and if the carrier is transmitted continuously it represents a considerable loss. In the case of 100 per cent modulation by a single-frequency tone, the carrier power is twice the sum of the power in the two sidebands.

Spread sidebands and carrier transmission uses sidebands displaced from their normal positions in relation to the carrier by an amount approximately equal to the audio band transmitted. It is sometimes used when inverter-type privacy is employed, because the spreading then can be accomplished without additional modulating equipment and merely requires that the inverter be designed for different input and output frequency bands. Advantage: less stringent distortion requirements are imposed on radio equipment because predominant intermodulation products fall outside the used band. Disadvantage: communication band, much wider than otherwise necessary, is occupied inefficiently.

Two sidebands, suppressed carrier, transmission has not been used, because the difficulties of correctly maintaining the phase of the reintroduced carrier at the receiving end generally offset the advantage of saving in power capacity at the transmitter. Furthermore, *suppression of one sideband as well as the carrier* usually offers additional advantages with less stringent synchronization requirements, since in this case the frequency of the reintroduced carrier can depart from the correct value by a few cycles without appreciable mutilation of speech quality, and it can be off as much as 20 cycles without noticeably reducing articulation.

Single sideband, suppressed carrier, transmission has certain important advantages which are especially valuable in long-wave systems where transmitter power capacity, communication band width, and static interference are controlling factors. These advantages, referred to double sideband and normal carrier modulated 100 per cent, are: 6-db increase in intelligence-bearing signal for same transmitter amplitude capacity, 3-db reduction in random received noise, no emission from transmitter during idle periods, and radio transmission band no greater than audio band width. Disadvantages: more complicated modulating process; precision frequency control is required in order to reintroduce at the receiver the carrier suppressed at the transmitter; transmitter power supply is subjected to large load variations at syllabic frequencies. This method is employed in the long-wave and short-wave systems between New York and London and on some radiotelephone systems. In the short-wave system, synchronization of the reintroduced carrier at the receiving end is accomplished satisfactorily by transmission of the carrier or an equivalent pilot signal reduced 10 to 20 db below normal carrier.

Privacy is usually achieved by modifying the signals in one or more ways which render the message substantially unintelligible unless received with special equipment. Devices called speech inverters have been developed which reverse the sequence of the audio frequencies before modulation in the radio transmitter; reinversion is then necessary after detection at the receiver. Additional privacy is obtained by varying the radio carrier in a cyclic manner. A more complicated method, known as split-band privacy, involves dividing the total audio band into several narrow bands. These can be inverted and/or transposed in various arrangements, and the combination can be changed frequently. All these methods have been applied successfully to radio circuits.

10. PRINCIPLES OF TWO-WAY OPERATION

SYSTEMS WITH FOUR-WIRE TERMINALS. Since the emission of radio waves and their subsequent detection at a distant point constitute inherently a unilateral process, duplex operation requires two one-way radio systems acting in opposite directions. Such an arrangement for telephony is shown in Fig. 6. At each end of the system a

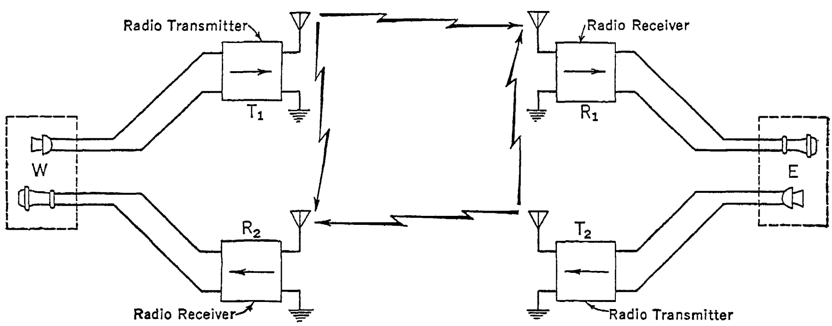


Fig. 6. Simple Radio System—Four-wire Terminals

microphone is connected to a radio transmitter and a telephone receiver is connected to a radio receiver. Thus, two persons can converse provided the radio transmission eastward does not interfere with the transmission westward; i.e., radiation from T_2 reaching R_1 must not prevent R_1 from functioning satisfactorily in receiving signals from T_1. Likewise radiation from T_1 at R_2 must not interfere with reception of signals from T_2. For the condition where the person at W is talking and the person at E is listening, the effect of T_2 at R_1 may be (*a*) increased noise, (*b*) distortion of signals from T_1, (*c*) detection of unwanted signals. Furthermore, the effect of T_1 at R_2 may result in a return signal to W which is quite disconcerting. If the person at W hears himself talking in reasonable volume, it will not disturb him unless there is sufficient delay to give an echo effect. However, if he hears unintelligible sounds in like volume, they will seriously disturb him. Such sounds appear if R_2 is overloaded by high field intensities from T_1 or if T_1 produces extra-band radiation within the selectivity band of R_2 and of sufficient intensity to be detected.

The forms of interference outlined above are prevented by adapting one or a combination of the following expedients: (*a*) use of different frequency bands for the two directions of transmission, (*b*) geographical separation of the radio transmitter and receiver, (*c*) use of directional antennas, (*d*) suppressing all emission from transmitters except during transmission of wanted signal and disabling and protecting local receivers during active intervals of adjacent transmitter. Accomplishing (*d*) involves switching operations. Manual control of the switching is satisfactory for experienced talkers, but voice-current control is favored for more general use. Single sideband, suppressed carrier systems are less susceptible to interference than other systems and may be used more readily without switching arrangements, if the transmitter and receiver are separated sufficiently.

In planning a specific system, there is a wide variety of circumstances, including economic factors and the characteristics of the particular apparatus to be used, which determine the selection of the method or combination adopted. There are no clearly defined dividing lines, but it may be stated rather generally that the use of (*a*) or (*d*) alone is satisfactory for short-distance working. All long-distance systems employ at least (*a*) and (*b*) or (*b*) and (*d*). Geographical separations range from a few hundred feet to several hundred miles. Use of (*c*) reduces the distance necessary. With (*a*) the necessary geographical separation is determined by the attenuation required to bring the ratio of field intensities of the wanted and unwanted signals within the limits of the receiver selectivity and overload characteristics. With (*d*), geographical separation is necessary for high-power transmitters, to avoid exposing the receiving system to noise due to spurious emanations from parts of the transmitting apparatus or the associated power-supply system.

SYSTEMS INTERCONNECTING WITH TWO-WIRE EXTENSIONS. Since standard telephone subscriber loops are two-wire circuits in which messages in opposite directions traverse the same wire path, the two oppositely directed radio paths in Fig. 6 must be arranged to terminate two-wire. The ordinary hybrid coil arrangement common in telephone repeaters and four-wire cable circuits fails to solve this problem except where the radio circuit meets all the requirements imposed by wire practice on corresponding wire circuits. This is seldom possible or economical on account of difficulties peculiar to the radio paths. In wire systems, transmission levels remain fixed within closely established limits, and signal volumes vary over a considerable range. In radio circuits, comparatively large variations in attenuation sometimes occur in relatively short intervals of time, except over extremely short-distance paths, and these tend to cause retransmission of received signals at such amplitudes that severe echoes and even singing around the two ends of the circuit will occur unless means are provided to prevent it. Furthermore, for all long-distance working, it is uneconomical to provide transmitter capacity which will permit appreciable variations in signal volume. To obtain maximum signal-noise ratios at the radio receivers, it is essential that the speech currents fully load the transmitters. This requires gain adjustments between the hybrid coils and the transmitters to suit the particular talkers and the condition in the connecting wire circuits.

To overcome these fundamental transmission difficulties, automatic switching systems operated by the voice currents of the speakers have been developed. These devices block the radio path in one direction while speech is traveling in the reverse direction and also keep one direction blocked when no speech is being transmitted. The operation is so rapid that it is unnoticed by the telephone users. Since these systems prevent the existence of singing and echo paths, their use permits the amplification to be varied at several points almost without regard to changes in other parts of the system, and it is possible by manual or automatic adjustment to maintain the volumes passing into the radio link at relatively constant values irrespective of the lengths of the connected wire circuits and the talking habits of the subscribers.

Figure 7 is a schematic diagram of one end of a circuit showing the essential features of a voice-operated device. This kind of apparatus is capable of taking many forms and is, of course, subject to change as improvements are developed. The diagram illustrates how one of these forms might be set up. This form employs electromechanical relays. The functioning of the apparatus illustrated is briefly as follows: the relay *TES* contact is normally open so that received signals pass through to the subscriber. The relay *SS* contact is normally closed to short-circuit the transmitting line. When the subscriber at *W* speaks, his voice currents go into both the transmitting detector and the transmitting delay circuits. The transmitting detector is a device that amplifies and rectifies the voice currents to produce currents suitable for operating the relays *TES* and *SS*, which thereupon short-circuit the receiving line and clear the short circuit from the transmitting line, respectively. The delay circuit is an artificial line through which the voice currents require a few hundredths of a second to pass so that when they emerge the path ahead of them has been cleared by the relay *SS*. When the subscriber at *W* has ceased speaking, the relays drop back to normal. The function of the receiving delay circuit, the receiving detector, and the relay *RES* is to protect the transmitting detector and relays against operation by echoes of received speech currents. Such echoes arise at irregularities in the two-wire portion of the connection and are reflected back to the input of the transmitting detector, where they are blocked by the relay *RES* which has closed and which hangs on for a brief interval to allow for echoes that may be considerably delayed. The gain control potentiometers, shown just preceding the transmitting and receiving amplifiers, are provided for the purpose of adjusting the amplification applied to outgoing and incoming signals.

The relief from severe requirements on stability of radio transmission and from varying speech load on the radio transmitters, which such a system provides, permits much greater freedom in the design of the two radio circuits than would otherwise be possible. In the system shown in Fig. 7 interference between local transmitter and receiver, as outlined previously in discussing Fig. 6, is prevented by such geographical separation as may be necessary in combination with either the use of two communication bands or single sideband, suppressed carrier, transmission. When one communication band is used for both directions with carrier and double sideband transmission, the switching systems of the type shown in Fig. 7 are extended to operate additional devices which suppress the carrier at the transmitter and disable the receiver. This switching is necessary to protect

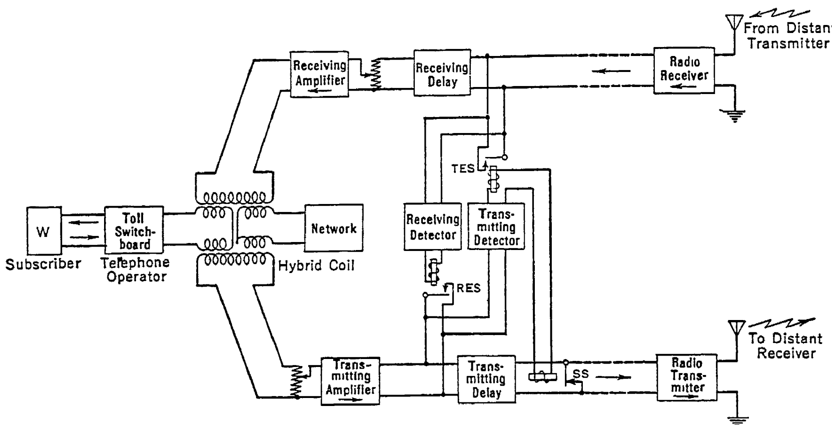


FIG. 7. Two-wire Radio Terminal Showing Arrangement of Voice-operated Switching Devices

the receiver where it is exposed to intense fields from the transmitter. When privacy apparatus is included in an installation, the voice-frequency switching is frequently arranged to transfer the same privacy unit from the transmitting to the receiving leg of the four-wire terminal (or vice versa), thus saving cost of a second privacy unit.

11. SYSTEM DESIGN

The designer of a radio circuit is limited to the establishment of certain facilities at the transmitting and receiving station which are expected to yield the desired results on the basis of available data, computations, and previous similar experience. Beyond this point, the performance is inherently a matter of probability, since transmission is subject to the vagaries of natural influences entirely beyond the control of man. Sometimes, very short-distance circuits can be provided which are substantially immune from these influences. However, the designers of long-distance circuits frequently encounter technical limitations which determine the maximum degree of reliability attainable, quite apart from considerations of cost.

The design proceeds from a statement, which includes (*a*) type and nature of service expected, (*b*) daily hours of operation, (*c*) general location of terminals, (*d*) distance covered and character of intervening region, (*e*) overall transmission requirements, (*f*) future plans. From these flow the requirements and compromises which ultimately determine the selection of transmission frequencies and methods; the location and general arrangement of transmitting, receiving, and voice terminal stations; and the choice of equipment.

Systems requiring separate transmitting and receiving stations usually involve expenditures which warrant a fairly comprehensive preliminary survey. The survey includes a search for suitable station sites and measurements of received field intensities, noise levels, and angles of wave arrival over the approximate path at substantially the proposed frequencies. This survey frequently can be effected by observing the signals from existing telephone or telegraph stations. It serves to substantiate conclusions derived from computations and from available transmission data; it gives important specific information concerning noise at receiving sites; and it should serve to disclose any conditions peculiar to particular locations which may have a profound effect upon performance.

TRANSMISSION REQUIREMENTS are related intimately with the needs and objectives of individual applications to such an extent that the statements herein should not be taken as concrete recommendations but should be regarded principally as guides to the items that need consideration when formulating the specific requirements for a particular system.

Circuits terminating four-wire and used only for direct conversations between terminals can be operated over a wide scale of conditions all of which may be acceptable for the purpose in hand. Circuits interconnecting with telephone plants should conform as far as possible to the standards of the wire systems in respect to transmission, stability, distortion, and interference effects. Allowance must be made for similar imperfect conditions in wire transmission. Otherwise the complete connection may be unsatisfactory, even though it is possible to converse over the radio circuit alone. With the possible exception of very short-distance radio circuits, it is generally true that radio transmission conditions vary through an extremely wide range, and it becomes necessary during some operating intervals to endure a comparatively poor circuit or temporarily do without one. It is desirable, therefore, to make the limiting requirements as liberal as possible.

Transmission Times. The requirements are no different from those for corresponding wire systems. The transmission time, from radio transmitter input terminals to receiver output terminals, is seldom appreciably greater than the time required for radio waves to traverse from transmitter to receiver. The propagation rate is approximately 186,000 miles per second. To this must be added time for wire circuits to the control centers, time for delay networks used with voice-operated switching devices, and time for any delays incurred in other apparatus, such as privacy systems.

Stability. It is desirable that all equipment adjustments that affect the overall circuit performance, and more particularly those that cannot be made without removing the circuit from service, retain the established conditions for long periods.

Receiving-set selectivity, in addition to discriminating sharply against unwanted signals, should provide ample margins for the total effect of frequency drifts to be tolerated over a period of several hours at all points in the system where base frequencies are generated. For example: a system employing a quartz-plate oscillator at the transmitter and a well-designed beating oscillator of the tuned circuit type at the receiver, when operating at 15 Mc and transmitting audio frequencies up to 3000 cycles, requires an intermediate-frequency band width of the order of 8000 cycles (-3 db at 4000 cycles from midband).

When the system is operated on the basis of constant net loss, it is desirable that the loss in the radio circuits be held to ± 1 db. When the system is operated on the basis of constant input volume to the radio transmitter and voice-operated devices are used, variations of ± 5 db are sometimes tolerated at the receiving end. When radio transmission is subject to rapid variations the receivers are provided with automatic volume controls which hold the output volume within a few decibels for wide fluctuations in the signal-field intensity.

Audio distortion is usually specified in terms of transmission band characteristics, amplitude or load characteristics, and departure from proportionality of phase shift with frequency. Difficulties from the last are seldom encountered in ordinary radio circuits used for voice communication.

Transmission Band Characteristics. The requirements in each instance are closely related to the needs of the particular application. Except when using radio frequencies less than 100 kc, it is seldom a difficult matter to provide reasonable bands for conversational purposes. The minimum requirements for international radio telephone circuits, as recommended by the International Consultative Committee on Telephony, are the same as for long-distance wire circuits: i.e., 300–2600 cycles as the limiting frequencies effectively transmitted. The committee also recommends that future circuits be designed to transmit at least 200–3000 cycles. Widening the band improves quality and to a lesser extent improves articulation, provided that other conditions do not become controlling. For instance, exposure to random noise at the radio receivers is substantially proportional to the band width, so that widening the band may not result in improved performance under adverse receiving conditions. In the United States, it is customary to specify the transmission-frequency characteristic in terms of departure in decibels from the transmission level at 1000 cycles.

Amplitude or Load Characteristics. Except when radio circuits are subject to rapidly varying transmission conditions, such as those frequently encountered in the short-wave range, load tests are made in the same manner as for wire circuits, and the requirements are the same for equivalent results.

The linearity requirements for the transmitting system and the receiving system are usually specified in terms of distortion products resulting from the application of pure tones. The tests are made separately for the two systems. Carrier and double sideband systems may be tested with a single tone at varying amplitudes. Single sideband systems require two-tone tests. If privacy apparatus is to be used with double sideband systems, it is usually required that the total distortion products introduced by the transmitter alone or the receiver alone and falling within the audio band shall be about 25 or 30 db below the single tone output for any tone input up to 90 per cent modulation. If two equal tones are used on any system, it is generally specified that any distortion product falling within the transmission band shall be 20 to 25 db below one tone. (For testing methods, see Section 11 and *I.R.E. Report of Standardization Committee.*)

Radio-frequency or phase modulation at the transmitters results in troublesome distortion effects at the receiving station if radio transmission occurs simultaneously over two or more paths differing in length by appreciable portions of the wavelength used. This situation is encountered in the short-wave systems, and it is sometimes specified that, during the modulation cycle, the phase shift associated with frequency modulation or phase modulation should not exceed $\pm 15°$. (For method of test, see *I.R.E. Report of Standardization Committee.*)

When no appreciable differences in simultaneous transmission path lengths are encountered, frequency modulation can be tolerated if all the radio equipment has substantially a flat frequency-transmission characteristic throughout a band sufficiently wide to pass all the essential frequencies generated and if extra-band radiation does not interfere with other services.

Interference is caused by signals from other radio circuits and by disturbances generally classified as noise.

Unwanted signals may enter the radio circuit through cross-modulation effects at the transmitting station if there are two or more transmitters, or they may enter at the receiver. Cross-modulation is likely to occur with open-wire transmission lines. It is not difficult to overcome if active lines are well separated and long parallel runs are avoided. It results from the impression on the tube circuits of one transmitter of modulated radio-frequency voltages generated by a second transmitter coupled to the first usually via the transmission lines or the antennas. In special cases, trap circuits or other simple filtering devices are introduced when found necessary, but they are objectionable if the same lines are to be used for several frequency assignments.

At the receiving station it is not enough to provide apparatus having comparatively high selective properties. It is necessary to know in what manner this selectivity is achieved, and what values of unwanted signal voltage may be impressed upon the receiver input-terminals simultaneously with the wanted signal. A receiving site and an antenna system must then be selected which will not violate these receiver requirements.

Noise at the receiving terminal is derived from the connecting wire system, the radio transmitting and receiving apparatus, and the radio noise field. If noise from the wire system meets the accepted standards for good toll circuits, as it should, it is not likely to have a noticeable effect on the performance of the radio circuit. Noise generated within the radio transmitter is measured in terms of audio signal by means of a linear monitoring rectifier exposed to the transmitter radio output. The audio signal-noise ratio thus obtained should be somewhat better than the maximum audio signal-noise ratio which it is desired to obtain at the receiving end under conditions of high signal-field and low noise-field intensities. Noise due to the receiving equipment should never be a controlling factor except when approaching the limit of sensitivity.

Since the effect of noise depends greatly on its frequencies in relation to the audio transmission band, precautions are necessary in systems employing frequency inversions to prevent the conversion of relatively harmless noise into very objectionable noise. This conversion may occur if the noise enters any part of the system between points where the inversions and reinversions are made.

The interfering effect of noise is very difficult to express accurately.

12. INSTALLATIONS

The successful establishment and maintenance of dependable, long-distance circuits with two-wire terminations require careful installation planning, the provision of adequate testing facilities, and the consideration of many problems only indirectly related to the technical operation of the system. There is a wide gap between this extreme and the simple facilities required for short-distance radio circuits without two-wire extensions (Section 7).

TRANSMITTERS AND RECEIVERS AT SAME LOCATION. Small transmitting systems for short-distance service are placed at the same location as the receiving systems. The transmitter and receiver are usually self-contained units requiring a single connection to the general power supply. If the same frequency assignment is used for both directions of transmission a single antenna is sufficient. Manual or voice-controlled switching is necessary to change from receive to transmit conditions. If two frequency assignments are used without quiescent transmitter operation, it is frequently found more satisfactory to employ two antennas slightly separated than to make provision for transmission and reception on the same antenna. The latter is possible by means of various special circuit arrangements but is almost certain to incur a penalty in respect to minimum receivable field intensities. Since the field intensity gradient around the transmitting antenna is extremely steep, it is seldom necessary to remove the receiving antenna more than 50 to 500 ft in order to secure satisfactory conditions. The distance depends on transmitter power, type of antenna, frequency difference, receiver selectivity, and load characteristics. If the antennas have directional properties, the relative positions should be selected, when possible, so that each antenna presents a null in the direction of the other. Usually the transmitting antenna is erected near the apparatus, and the receiving antenna is placed at a distance from the receiver, connections being provided by suitable transmission lines.

If it is essential that the transmitting and receiving apparatus be installed in close proximity, attention needs to be given to shielding to prevent direct interference between transmitter and receiver. Receivers designed for this type of installation seldom require further shielding when used with transmitters up to about 25-watt capacity, provided that the transmitters are also reasonably well shielded. It is a good plan to place receivers somewhat away from transmitters for ratings up to about 500 watts. The alternative is to provide a special shielded compartment for the receiver. This has been done on shipboard, where space limitation and operating convenience demand a compact installation.

As the transmitter power is increased, the possibilities increase rapidly that noise will enter the receiver directly or through the receiving antenna from various sources within the transmitter or its power circuits. If this occurs, it limits the permissible receiver sensitivity and may completely nullify the value of higher power for the purpose of working greater distances. Recourse to voice-controlled quiescent transmitter operation greatly alleviates this type of interference for installations where two or more transmitters are not to be used simultaneously. It is frequently applied to ship systems and materially increases the working distances. It is effective only in eliminating noises related to the suppressed radio signal components. Around large transmitters the residual noise after the carrier or other radio signal components are suppressed still prevents the use of extremely sensitive receivers. This is one of the compelling reasons for establishing separate transmitting and receiving stations for long-distance circuits, where extremes of power and sensitivity are essential.

SELECTION OF TRANSMITTING STATION SITE. Items requiring consideration are: ground conductivity and dielectric constant; general character of surrounding terrain; position relative to receiving stations; possibilities of interfering with broadcast reception or that of other services; transportation, power, and telephone facilities; living arrangements for station personnel; prevalence of sleet storms; unusual conditions of temperature, humidity, presence of salt spray; etc. Ground conditions affect antenna design. Desirable characteristics depend upon the type of radiating system to be employed, the frequencies and the wave angles of transmission (Sections 6 and 10). The terrain in the direction of transmission affects the vertical wave angle. Mountains and hills subtending large angles are undesirable. Steel towers, buildings, transmission lines, etc., constituting sizable obstructions directly in front of short-wave directional antennas, are objectionable since they modify the directive pattern.

THE TRANSMITTING STATION LAYOUT is based primarily on the requirements of the antennas and their relation to the transmitter. Usually the antenna or antennas are located a short distance away from the building housing the transmitter, and connection is made through open-wire or concentric transmission lines. The practice of bringing the antenna downlead directly to the transmitter is seldom followed in modern installations. Use of uniform transmission lines, all having the same impedance, greatly simplifies switching problems. Good values are: open-wire 600 ohms, concentric lines 70 to 80 ohms. Placing the antennas clear of the transmitter building avoids difficulties in erection and maintenance. Short-wave directional antennas are placed so that the building is not within the horizontal angle of the principal lobe.

RECEIVING STATION EQUIPMENT depends somewhat on the number of radio circuits involved but is also influenced considerably by the standards established in respect to service interruptions other than those attributable to the transmitting medium. The essential components are: radio receiver and its power-supply units, antennas and transmission lines, wire-terminal apparatus and voice-frequency testing equipment with associated power-supply units, general power supply, including emergency power sources. Large stations usually have facilities for observing the field intensities of received signals and noise and for the precise measurement of received frequencies. These are used to check the transmitters which are a part of the system, and those of other systems, which create interference.

SELECTION OF RECEIVING STATION SITE is a matter demanding careful consideration, especially if long-distance services are contemplated, and it is seldom safe to make a final decision without actual observations of signal-field and noise-field intensities over a period sufficient to obtain representative data. Items that should receive attention are: location relative to transmitting stations of same system and all other nearby transmitters and sources of man-made noise; local ground conditions and general character of surrounding terrain in the direction of wave arrival; transportation, power, and telephone facilities; living arrangements for station personnel; prevalence of electrical storms; unusual conditions of temperature, humidity, presence of salt spray, etc.

It is desirable to have the antennas present nulls toward all transmitters in the area which are likely to produce any form of interference.

Likely sources of man-made interference are: high-tension transmission lines, electrical machinery in factories, electrical trains, automobiles, airplanes, motorboats, etc. The last three mentioned are particularly important in long-distance short-wave reception at times when signal-field intensities are low as the result of magnetic disturbances. With an extremely sensitive receiver and a directional antenna designed for low-angle reception, no serious interference would be expected from automobiles 1 $^1/_2$ miles in front of antenna and $^1/_2$ to $^1/_4$ mile at the sides and rear.

Reception of short waves arriving at low angles can frequently be improved from 5 to 10 db by placing properly designed antennas on ground sloping uniformly downward in the direction of the transmitting station at an angle of 5° to 15°.

RECEIVING STATION LAYOUT. Primary objectives are to place the antennas in an advantageous position for the collection of energy from the incoming waves and to obtain an efficient arrangement in respect to transmission lines. In choosing directional antenna locations, close attention should be given to the position of objects capable of reflecting or otherwise redirecting unwanted waves into the sensitive angles of the antenna characteristic.

It is well to adopt a uniform impedance for all transmission lines. Convenient values are: open wire 600 ohms, concentric lines 70 to 80 ohms. In order to avoid disturbing the incoming radio waves, and also to avoid undesirable currents, transmission lines should be placed as near the ground as practicable. However, it is inadvisable to place two-wire lines less than 6 ft from the ground. Four-wire balanced lines are disturbed less by the proximity of the ground and have been used successfully at 4-ft elevations. Concentric lines may be installed underground. They should always be sealed and, in some situations, are further protected from moisture by maintenance of pressure with inert gas. Aside from somewhat higher first cost, much can be said in favor of concentric conductors, since they substantially eliminate the difficulties encountered with converging lines at receiving set locations. At short-wave stations, it is desirable to have all directional antennas present a null to the building and the road approaching it.

SHIP STATIONS are usually installed in extremely limited quarters. They require compact units, designed to allow inspection and repairs without having access to all sides and preferably without disconnection and removal of parts. Rapid frequency-changing features are especially important if it is desired to maintain close contact with more than one shore station.

Transmitting and receiving antennas are usually separated as much as possible. Simple types are generally used because they are suitable for several frequencies. Horizontal directional properties are not desirable. Electrical noise conditions vary widely with positions aboard ship.

A four-wire termination is usually employed to avoid rather expensive control-office equipment. Ship's passengers talk from a conveniently located booth. Circuits are also provided from the captain's quarters or a similar point convenient to the bridge. With a four-wire terminal, no voice-operated switching apparatus is needed other than the simple devices necessary for quiescent transmitter operation. Voice-frequency apparatus is mounted adjacent to the receiver. One attendant supervises all operations and performs the duties of a technical operator. If the ship is equipped with radio telegraph as well as telephone apparatus, and the two must operate simultaneously, precautions are necessary to avoid mutual interference.

SHORE STATIONS are usually equipped to offset, as far as practicable, unfavorable conditions aboard ship. This is done by providing transmitter power capacity from 10 to 40 times that of the ship station, by employing directional antennas, and by selecting a quiet receiving location a few miles away from the transmitting station. Antennas having moderately directional patterns (6 to 12 db maximum net gain) are used to cover the principal ship lanes effectively. Less directional antennas are needed for general coverage. When a ship is close in, at which time the required direction of transmission is likely to move rapidly through a wide horizontal angle, antenna gain is less important, and a directional antenna, unless it has pronounced nulls, is frequently found satisfactory because the unfavorable ship position is offset by the short distance. Otherwise, a simple antenna is provided for this purpose.

BIBLIOGRAPHY

Articles in *Proc. I.R.E; Elec. Rev.* (London); *Elec. Communication; Bell Syst. Tech. J.; Trans. A.I.E.E.; Elect. Engg.; Proc. I.E.E.* (Engl.); *Electrician; Wireless World; Bell Lab. Rec.; Rev. gén. del'élec.;* and *Ann. des pastes.*

TELEPHONE LINES—TRANSMISSION CONSIDERATIONS

13. TYPES OF PLANT

SUBSCRIBER LINES (LOOPS) include all types of outside plant facilities needed to connect the subscriber station telephones with their local central office. Such facilities may consist of cable, either aerial or underground, open (bare) wire, carrier or radio channels, or a combination thereof. Generally, cable is used in urban areas as a means of serving local subscriber stations; where small urban open-wire plants are still in operation they are rapidly being replaced. In rural areas, open-wire lines are still in general use except in congested sections, since the fewer lines and longer distances characteristic of rural areas make the open wire more economical. As stated above, however, it is anticipated that radio facilities can be made available, commercially, to serve distant or relatively inaccessible farms *where the costs of providing the usual telephone wire facilities would be excessive.*

TOLL LINES, as generally defined, consist of various types of outside plant facilities employed to provide toll circuits between *toll centers* (*TC*). Those line facilities connecting TC and *tributary offices* are considered part of the TC plant. These latter offices, in the general meaning of the word "tributary," are small offices (in territory adjacent to the TC) connected to the TC by one or more tributary circuits and are fully or partly dependent upon the TC for the handling of their toll traffic. Toll facilities may consist of cable, aerial or underground, open wire, carrier or radio channels, or a combination of these.

Subscriber line facilities, known generally as *exchange* plant, and toll line facilities should be designed and constructed to meet, cooperatively, the overall service objectives which are known collectively as *service standards.* These standards are not fixed for all time but change with service needs and advancements in the art of communications.

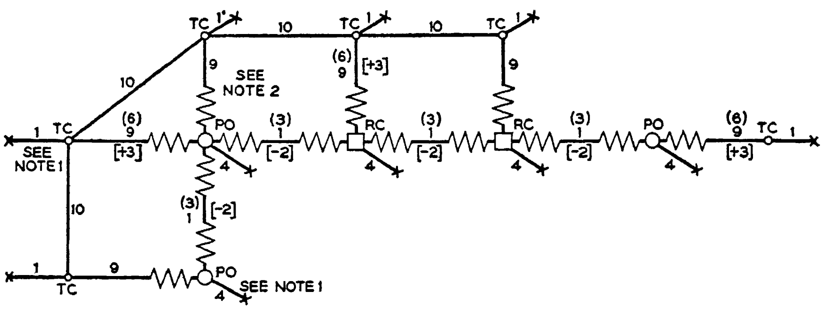
TRANSMISSION AND SIGNALING are two fundamental factors to consider in any telephone system, whether the connecting facilities between subscribers are wire, carrier, radio, or a combination of two or more of these types. If either transmission or signaling, or both, are not satisfactory for a given telephone system, the system is not workable under modern standards of service.

14. SERVICE REQUIREMENTS—TOLL

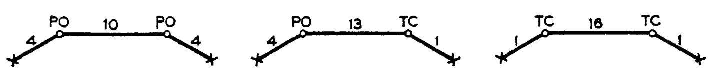
UNIVERSAL SERVICE is the goal toward which the telephone industry has been striving for many years and which, it now appears, will be attained. This goal simply means that anyone, anywhere, can talk, telephonically, with anyone else, anywhere else, whether the connection be established locally, within the nation, or between any two countries in the world. For a number of years it has been possible to talk by telephone from any point in the United States to any other point connected to the nationwide toll system and to many foreign countries. Worldwide service is being rapidly expanded to include those countries not at present reached.

TRANSMISSION OF SPEECH between two points requires that speech (sound) power from the talker actuate his transmitter diaphragm and that the transmitter convert this power into electrical power, which travels to the distant listener's receiver, where it is reconverted by the receiver diaphragm into speech (sound) power of approximately the same characteristics as the original speech power. It is obvious that the electrical power will diminish as it travels over the circuit between the talker and listener stations as the result of series and shunt impedances encountered in the lines and equipment which constitute the circuit. If the electrical power reaching the receiver is diminished to the point where it no longer drives the receiver diaphragm sufficiently to permit the listener's ear to interpret the intelligence carried by the resulting sound waves, then the electrical transmission loss in the circuit is too high to permit carrying on a satisfactory telephone conversation. It, therefore, becomes necessary to limit overall transmission losses and consequently the losses in component parts of circuits, which usually consist of two or more sections or links.

THE GENERAL TOLL SWITCHING PLAN (Fig. 1) as developed for establishing toll connections on a manual basis provides a practicable plan for accomplishing universal



FOR CIRCUITS DESIGNED TO HANDLE ONLY TERMINAL BUSINESS THE WORKING NET LOSS DEPENDS UPON TOLL TERMINAL LOSSES INVOLVED AND UPON OVERALL DIRECT STANDARDS. AS A TYPICAL EXAMPLE, WITH A DIRECT STANDARD OF 18 DECIBELS, THE LIMITING VALUES OF TERMINAL CIRCUIT NET LOSSES WOULD BE AS SHOWN BELOW FOR THE ASSUMED VALUES OF TOLL TERMINAL LOSS.



NOTES:

1. THE TOLL TERMINAL LOSS WOULD BE DETERMINED IN EACH CASE ON THE BASIS OF MEETING THE TRANSMISSION REQUIREMENTS IN THE MOST ECONOMICAL AND SATISFACTORY MANNER, FOR EXAMPLE, IN MEETING THE 10 DECIBEL LIMIT FOR OUTLET TERMINAL LOSS. THE VALUES SHOWN ABOVE ARE TYPICAL BUT THE ECONOMICAL TOLL TERMINAL LOSS IS EXPECTED TO VARY CONSIDERABLY IN INDIVIDUAL CASES.
2. PAD VALUES DEPEND ON NOISE AND CROSSTALK CONDITIONS, ON LIMITING TOLL TERMINAL LOSSES AND ALSO, IN THE CASE OF INTERMEDIATE LINKS, ON ECHO MARGIN REQUIREMENTS.

STANDARDS	DECIBELS
DIRECT	17-20
2-LINK THRU PO	20
2-LINK THRU TC	22
3-LINK	21
4-LINK	22
5-LINK	23
OUTLET TERMINAL LOSS	10

CODE

TC	TERMINATING TOLL CENTER
PO	PRIMARY OUTLET
RC	REGIONAL CENTER
-ꟿ-	SWITCHING PAD
x—1—	INDICATES TYPICAL TOLL TERMINAL LOSS
(6)	MINIMUM WORKING ECHO NET LOSS (ASSUMES NO TRANSMISSION IMPAIRMENTS)
9	EFFECTIVE WORKING NET LOSS-VIA CONDITION
[+3]	ECHO MARGIN

FIG. 1. Typical Example of the General Toll Switching Plan, Showing Limiting Toll Circuit Losses (Courtesy Bell System)

service in the United States and throughout the world, with the development and expansion of telephone systems in other countries.

From Table 1 it will be noted that any two subscriber telephones which have access to the nationwide toll network can be connected together, using not more than five toll circuit links and four switches. It is assumed that the telephones are located at their respective toll center (TC) points, i.e., not at tributary points, which would necessitate using a tributary trunk to reach the respective TC office.

The plan provides for *eight regional centers* (*RC*), Atlanta, Chicago, Dallas, Denver, Los Angeles, New York, San Francisco, and St. Louis, each being strategically located within the United States, to serve as toll switching centers of the first order. Each of these centers is connected by direct, high-grade toll circuits to each of the other centers. Within each RC area are a number of important toll centers known as *primary outlets* (*PO*), each being connected by direct, high-grade toll circuits to its own RC, other RCs and other POs, as required to best handle the traffic. Finally, each PO serves directly all the *toll centers* (*TC*) within its area, and the TCs serve the subscribers within their local areas, either directly or through their tributary offices. Thus, the plan provides for a concentration of the toll traffic at the various toll centers which have access to or are accessible from any part of the nationwide system (including Bell and Independent) through direct or switched connections.

Table 1 shows overall standards and number of links and switches, and Table 2 shows assigned losses for the different toll links, under the present general toll switching plan.

Table 1. Overall Standards

Toll Connection * (for switched traffic)	Overall Standard, decibels	Number of Circuit Links	Number of of Switches
Direct	17–20	1	0
TC-TC-TC	22	2	1
TC-PO-PO-TC	21	3	2
TC-PO-RC-TC	21	3	2
TC-PO-RC-RC-TC	22	4	3
TC-PO-RC-RC-PO-TC	23	5	4

* The letter codes are defined in Fig. 1.

In general, four-wire circuits (a separate path for each direction of transmission) or carrier channels are employed for the long-haul, intermediate toll links because of their better performance at low losses than that of two-wire circuits. Two-wire circuits are generally used for the shorter end links and toll trunks.

Table 2. Allowable Toll Link (Circuit) Losses

Toll Link (for switched traffic)	Effective Working Net Loss, decibels	Minimum Working Echo Net Loss, decibels (assumes no transmission impairments)	Echo Margin, decibels
TC-PO (end link)	10-(TTL) *	7-(TTL)	+3
PO-PO (intermediate link)	1	3	−2
PO-RC (intermediate link)	1	3	−2
RC-RC (intermediate link)	1	3	−2
PO terminal loss	10 †		

* This value depends upon the most economical and practicable toll terminal loss for each individual toll center which will meet the required *primary outlet* (*PO*) *terminal loss* of 10 db.

† This value results from taking one-half of the loss (20 db) for a two-link connection through a gain center (PO or RC).

The PO terminal loss of 10 db is fixed, unless changed under the plan. This loss may be allocated to the TC-PO circuit and the toll terminal loss as required. Toll terminal losses (TTL) vary from about 0 to 5 db.

TOLL CIRCUIT OPERATING REQUIREMENTS. The *minimum working net loss* (*MWNL*) of a toll circuit is the lowest net loss that may be assigned that will satisfy the design objectives imposed by singing, echo, crosstalk, and noise, when subject to maximum negative transmission variations.

The minimum working echo net loss (MWENL) is the lowest 1000-cycle net loss which can be assigned so that a circuit will satisfy the echo objectives, including the assigned echo margin. If the loss at which a circuit is operated is greater than the loss required to offset the echoes arising from the circulating current paths within the circuit, a *positive echo margin* is said to result. If the reverse is the case, a *negative margin* will be introduced. In order to operate the plant at lowest over-all losses on built-up or switched connections consisting of two or more links, positive echo margins have been assigned to some classes of circuits and negative margins have been assigned to others. Switching arrangements are so designed that negative margins will be offset in any connection.

Echoes are the result of imperfect balances in toll circuits equipped with telephone repeaters and four-wire terminating sets. The two-wire circuits are necessarily converted to four-wire circuits at each repeater, and the four-wire circuits require four-wire terminating sets at their terminals to convert the four-wire circuit to two-wire before extending it to the toll switchboard. At certain important switching offices, four-wire switching on a mechanical basis may be applied. At the points of conversion, a balancing network terminates one branch of the hybrid coil, and the opposite branch of this coil is connected to the outgoing or incoming toll circuit or the extension to the switchboard. It is not practicable to match the impedance of the toll circuit exactly or the extension with the impedance of the network. Thus, part of the voice currents are transferred across the hybrid bridge into the repeater inputs in varying degrees at each repeater or terminating set on the circuit instead of dividing equally between the outgoing line and its balancing network. These currents which enter the repeater inputs will be amplified and travel back to the talker with some delay, so that he hears his own words (in reduced volume) coming back to him an instant after he has spoken them. The listener may also be affected. As the delay and return volume increase, the annoyance becomes greater and may reach the point at which the conversation is not satisfactory. Figure 2 shows minimum working net losses for terminal grade circuits as limited by echoes, versus typical lengths of toll cable without echo suppressors and with anti-sidetone sets.

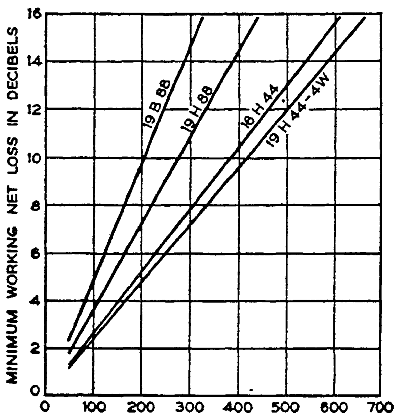


FIG. 2. Minimum Working Net Loss of Terminal Grade Circuits as Limited by Echoes, versus Typical Toll Cable Circuit Lengths without Echo Suppressors and with Anti-sidetone Sets (Courtesy Bell System)

To overcome echo difficulties, repeater gains must be properly assigned and regulated. Also, an echo suppressor has been developed which, by voice current action, causes a high loss to be introduced in one side of the four-wire circuit while speech is being transmitted on the other side. A simple schematic of this device is shown in Fig. 3. Generally, two-wire circuits are relatively short and do not require echo suppression.

The minimum working singing net loss (MWSNL) is the lowest net loss assignable to a circuit which will meet singing requirements. As indicated above under "echoes," the unbalances existing in practice at the hybrid coils of telephone repeaters and four-wire terminating sets cause part of the outgoing energy from one branch of the four-wire circuit to pass through the hybrid coil bridge points to the opposite transmitting branch of the four-wire circuit. At each hybrid coil this action occurs, so that, whether the circuit consists of only a two-wire repeater with terminating lines or a long four-wire circuit with several intermediate repeaters and a four-wire terminating set at each end of the circuit, a circulating current is established, provided that the net circuit gains exceed the net circuit losses in the circulating current circuit and the phase change in the circulating current is a multiple of 360°.

It is thus necessary in designing circuits, particularly two-wire, to limit these circulating currents by assigning repeater gains, so that, for an average circuit net loss, the most critical two-wire repeater for 95 per cent of the connections will have losses which total at least 10 db more than the gains in the two directions of transmission. For the short terminal circuits with one or two repeaters, an 8-db *singing margin* will usually be satisfactory.

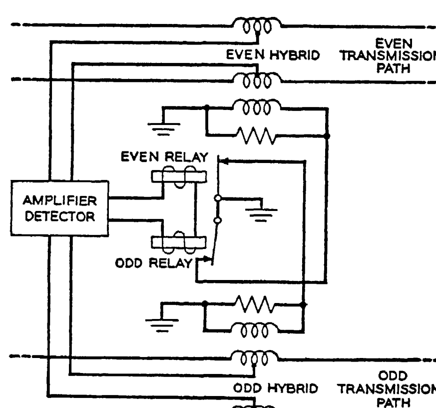


FIG. 3. Schematic of Echo Suppressor Circuit (Courtesy Bell System)

Figure 4 shows minimum working net losses, as limited by singing, versus typical toll cable circuit lengths for specified conditions of repeater spacing and singing points.

The minimum working crosstalk net loss (MWXNL) is the lowest net loss assignable to a circuit which will satisfy crosstalk requirements under all operating conditions. *Crosstalk* is the electric and magnetic transfer of speech or similar currents from one telephone message circuit to another. It may or may not be intelligible, but when it is composed of confused noise from several sources it is known as *babble*.

Crosstalk usually results from cumulative, slight unbalances between circuits or high energy level differences acting through close couplings in cable or open wire. The transferred energy is amplified wherever telephone repeaters are present.

Two types of crosstalk, *near end* and *far end*, develop between circuits. The first type travels to a listener on one circuit in a *direction opposite* to the transmission from a talker on another circuit; the latter type travels to the listener on one circuit in the *same direction* as the transmission from a talker on another circuit. Near-end crosstalk occurs in wire but not in properly arranged carrier circuits (except W.E. Co. G-type); far-end crosstalk appears in both wire and carrier circuits.

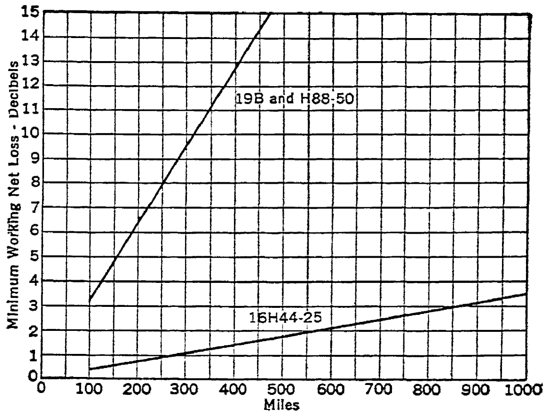


FIG. 4. Minimum Working Net Loss as Limited by Singing, versus Typical Toll Cable Circuit Lengths (Courtesy Bell System)

The effect of crosstalk on subscriber conversations depends not only on the actual volume of crosstalk heard but also on circuit and room noise present, circuit losses, and personal reactions.

Crosstalk is controlled primarily by avoiding excessive energy level differences and couplings[1] between adjacent parallel circuits. Techniques have been developed to limit both level differences (by proper regulation of repeaters and other amplifiers) and high couplings and to employ different frequency bands in controlling crosstalk.

Figure 5 shows minimum working net losses, as limited by crosstalk, versus typical *toll* cable circuit lengths.

Crosstalk values have generally been expressed in terms of *crosstalk units*, which are defined as one million (10^6) times the ratio of the crosstalk current or voltage at the observing point on the *disturbed circuit* to the current or voltage at the sending point on the *disturbing circuit* (assuming equal impedances at these two points). If the impedances are not equal, the square root of the power ratio may be used in place of the current or voltage ratio. With the development of visual indicating apparatus for measuring crosstalk and noise, crosstalk measurements have more generally been made in terms of *crosstalk coupling loss-db*, which means the net transmission loss between the sending point on the disturbing circuit and the receiving point on the disturbed circuit, it being understood that the higher the measuring set reading (loss), the less the actual coupling. More recently, the term *db above reference coupling-dbx* has come into use. This term means the coupling in decibels above reference coupling, and reference coupling means the coupling which would be required to give a reading of 0 dba on a W.E. Co. 2B noise-measuring set connected to the disturbed circuit when a test tone of 90 dba (using the same weighting as on the disturbed circuit) is impressed on the disturbing circuit.

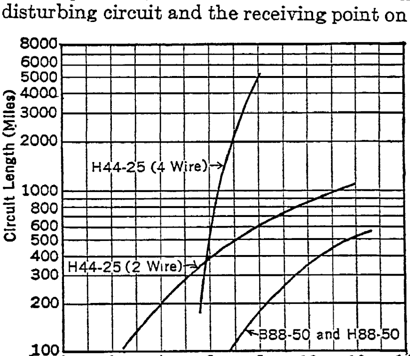


FIG. 5. Minimum Working Net Loss as Limited by Crosstalk, versus Typical Toll Cable Circuit Lengths (Courtesy Bell System)

The 2B set is designed to measure crosstalk and noise volumes or couplings (as well as other quantities) in decibel values. These values can be adjusted to a common basis for different types of lines and telephone receivers so that a given adjusted value, designated *dba*, will mean the same interfering effect to the ear, regardless of the type of line or subscriber set, affected by the crosstalk or noise being measured.

Figure 6 shows the relation between the terms crosstalk units (cu), crosstalk coupling loss, db, and crosstalk coupling, dbx.

The present design requirements for crosstalk limitations in circuits are taking into account the wide reactions of different people to different amounts of crosstalk, the variation in crosstalk volumes due to the variation in speech power of different talkers, the action of room and line noise on crosstalk effects, the intelligibility of crosstalk, costs involved in its control, and attainment of a good balance in judging the importance of the various factors entering into circuit design. Present practices indicate that it is inadvisable to permit crosstalk couplings in excess of 30 dbx (equivalent to 60 db crosstalk coupling loss or to 1000 crosstalk units), as measured from the transmitting to the receiving switchboard for a single disturber circuit.

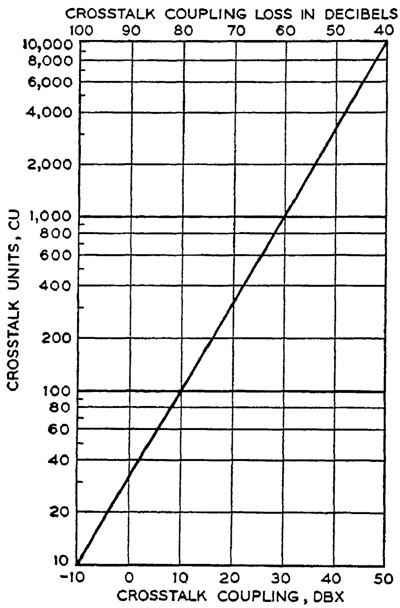


Fig. 6. Chart showing Relation between Crosstalk Units, cu; Crosstalk Coupling Loss in Decibels; and Decibels above Reference Coupling, dbx (Courtesy Bell System)

Noise transmission impairments which may exist in toll circuits due to power line induction, noise generated within telephone systems, or circuit irregularities must be limited to avoid transmission penalties in circuit operation. Noise reaching the subscriber's ear through his telephone receiver is, in fact, equivalent to adding loss to the circuit. Table 3 shows how these penalties are evaluated in terms of *noise levels-dba.*

Reference noise (RN) is used as a base in the calculation of circuit noise in terms of decibel penalties, as given in Table 3. Reference noise is registered as 0 dba on a 2B noise-measuring set when the input into this set is 10^{-12} watt of 1000-cycle power (line weighting). RN is equivalent, when measured at the terminals of a 600-ohm line (with line weighting), to 7 noise units; if measured across the terminals of a W.E. Co. No. 144 receiver, it is equivalent to 14 noise units.

Circuit noise of 29 db above RN (200 noise units) or less in a 600-ohm line is not considered to offer any appreciable noise impairment to a conversation, but for about each 3-db increase in noise level above 29 db the impairment increases 1 db, which must be included in the overall circuit loss.

Remedial measures have been perfected for controlling most types of noise to avoid penalties.

Distortion transmission impairment (DTI) to conversations results from a restricted or modified transmission of the full voice-frequency band necessary for clear, understandable speech. Such restriction or modification may be due to a low cutoff frequency of certain types of loaded line facilities and line apparatus. The older H172-63 loaded cable facilities and certain early types of telephone repeaters, carrier systems, and filters give distortion impairments. The latest types of loaded cable facilities, such as H44-25 and H and B88-50, and the latest-type standard repeater and carrier systems are considered to offer no appreciable distortion for the usual lengths employed. Figure 7 and Table 4 show distortion impairments for different facilities with FIA-AST subscriber sets and H88 switching trunks, as used in the Bell System.

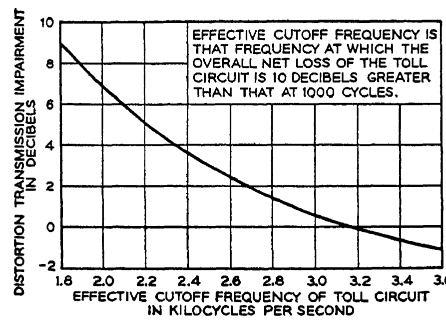


Fig. 7. Diagram Showing Distortion Transmission Impairment for Different Facilities with FIA-AST Subscriber Set and H-88 Switching Trunk, versus Effective Cutoff Frequency of Toll Circuit (Courtesy Bell System)

Volume transmission losses in toll circuits will vary with changes in temperature and, in addition, for open-wire facilities with such conditions as rain, sleet, and snow. These transmission variations are different for aerial and underground cable and open wire and also change with frequency.

Table 3. Noise Transmission Impairments Corresponding to Noise Magnitudes Measured or Computed at Various Points in the Transmission Circuit

Type of circuit		Toll circuits	Toll connecting, tandem and interoffice trunks	Local loops	
Point of measurement or estimate		Receiving toll switchboards	Local office	Subset terminals of loop	Across receiver terminals
Impedance of Measuring device		600 ohms, terminating, 6000 ohms bridging *	600 ohms, terminating, 6000 ohms bridging *	600 ohms terminating	Approximately 2000 ohms
Noise Rating	Noise Transmission Impairment, decibels	Noise Magnitudes in dba			
		Line Weighting	Line Weighting	Line Weighting	Receiver Weighting
N0	0	0 –29	0 –26	0 –20	0–17
N1	1	29.1–32	26.1–29	20.1–23	17.1–20
N2	2	32.1–35	29.1–32	23.1–26	20.1–23
N3	3	35.1–38	32.1–35	26.1–29	23.1–26
N4	4	38.1–40	35.1–37	29.1–31	26.1–28
N5	5	40.1–42	37.1–39	31.1–33	28.1–30
N6	6	42.1–43	39.1–40	33.1–34	30.1–31
N7	7	Over 43	Over 40	Over 34	Over 31

* When bridging measurements are made on a line, 600-ohm impedance is assumed each way from the bridging point. Under these conditions (*a*) with *144 rec.* line weighting, a correction † of +11 db is added to 2B set readings and (*b*) with FIA line weighting, a correction † of about +18 db is added to 2B set readings, in order to express readings in dba. If impedances each way from bridging point are not 600 ohms, correct 2B set readings as follows:
Each impedance 300 ohms, correction is +3 db.
Each impedance 400 ohms, correction is +1.8 db.
Each impedance 900 ohms, correction is −1.8 db.
Each impedance 1200 ohms, correction is −3.0 db.
Each impedance 2000 ohms, correction is −5.2 db.
† The corrections thus indicated should be added before entering Table 3.

Present Bell System practices provide for maintaining toll circuits of minimum to maximum lengths within the following deviations from the specified value: voice-frequency (VF) cable, ±1.0 to ±4.0 db; K-carrier, ±2.0 db; VF open wire, ±1.0 to ±3.0 db; open-wire carrier, ±2.0 db; and various combinations and lengths of facilities, ±1.0 to ±4.5 db.

Such limitation is provided for long VF cable circuits by devices known as *automatic transmission regulators* spaced at proper intervals along the circuit to automatically add or reduce gain in the circuit as required to maintain the specified volume limits. Figure 8 shows a pilot wire transmission regulator circuit, with its pilot wire cable pair in the same cable as the regulated cable circuits. Long cable circuits have a regulating repeater (in place of the regular repeater) about every 150 miles. These repeaters are controlled at each point by a master regulator, in accordance with temperature change in the pilot wire, which causes the repeater gains to vary above or below normal setting in steps varying from 1/4 to 1 db over a range varying from 2.75 to 19 db as required to maintain normal level. No system of automatic regulation has seemed necessary for use with open-wire voice-frequency facilities.

Since in cable circuits the attenuation-frequency curve is appreciably different at different temperatures, it is necessary to correct for this difference, known as *twist*, for high-frequency carrier systems, such as the K-system. The twist effect in a 100-mile aerial toll cable is shown in Fig. 9. Twist correcting circuits, as shown in Fig. 10, are located in long cable circuits about every 100 miles for aerial and 200 miles for underground cable.

For open-wire carrier systems, pilot channel regulator equipment is incorporated in the carrier terminal and repeater design to maintain transmission levels.

The effective transmission loss of a toll circuit is equal to the *1000-cycle loss* plus any *noise or distortion transmission impairments*, all expressed in decibels.

The overall effective equivalent of a complete toll connection from subscriber to subscriber is the sum of the *effective transmission loss* of the toll circuit or circuits and the *toll terminal losses* at the terminating toll centers.

Table 4. Distortion Impairments

Voice Frequency Facility	2-Wire or 4-Wire	Filter (Note 4)	Length in Miles for Various Impairments							
			−1 db	0 db	+1 db	+2 db	+3 db	+4 db	+5 db	+6 db
H245-S	2-W	No repeater					0–10	10–40	40–70	70–100
	2-W	A						0–7	7–30	30–50
H155-P	2-W	No repeater					0–20	20–60	60–100	
	2-W	A						0–10	10–40	40–50
H174-S, H172-S	2-W	No repeater			0–35	35–70	70–100			
	2-W	B			0–15	15–50	50–90	90–160	160–270	
	4-W	B					0–300			
	4-W	A							Over 300	
H106-P	2-W	No repeater			0–60.	60–110				
	2-W	B			0–25	25–60	60–150	150–220	220–300	
	4-W	B					0–300			
	4-W	A							Over 300	
H63-P	2-W	No repeater	Any							
	2-W	C			0–75	75–180	180–450			
	4-W	C		0–300	300–700					
	4-W	B				Over 700				
B or H88-50, S or P	2-W	No repeater	Any							
	2-W	D	0–150	150–450						
	2-W	C			0–100	100–250	250–400	Over 400		
	4-W	D	0–200							
	4-W	C			Any					
H44-25, S or P	2-W	No repeater	Any							
	2-W	D	0–150	150–450						
	2-W	C			0–100	100–320	Over 320			
	4-W	D	0–800							
	4-W	C		0–1000	Over 1000					
N.L. open-wire side with 3KC carrier line filters	2-W	No repeater		Any						
	2-W	1059B			0–800	Over 800				
	2-W	C			0–300	Over 300				
N.L. open-wire phan. or sides with 5 kc or no carrier line filters	2-W	No repeater	Any							
	2-W	D	Any							
	2-W	1059B			0–800	Over 800				
	2-W	C			0–300	Over 300				
Type of Carrier Frequency Circuit			Maximum Number of Links for Various Impairments							
			−1 db	0 db	+1 db	+2 db	+3 db	+4 db		+9 db
C2, C3, C4—Manual regulation						1	2	5		
C2, C3—Automatic regulation					1	2	4	Over 4		
C4—Automatic regulation					2	5				
C5				Any						
D						1	2			
EB										Any
G1			2							
H1—No repeaters			1	3						
J			Any							
K			Any							
L			Any							

Notes:

1. Impairments are referred to a distortionless toll circuit containing a 250–3000 cycle band-pass filter having square cutoffs.
2. Impairments are substantially independent of gage.
3. Impairments are substantially independent of type of line repeater, provided standard equalization is employed.
4. A refers to 13A or 32A filters; B, to 13B, 32B or 128B filters; C, to 13C, 32C or 128C filters; D, to D93985, 32C modified per KS-4165 (D160523), or 128A filters. 1059B filter is associated with the high level 22-type repeater (104-D tubes).

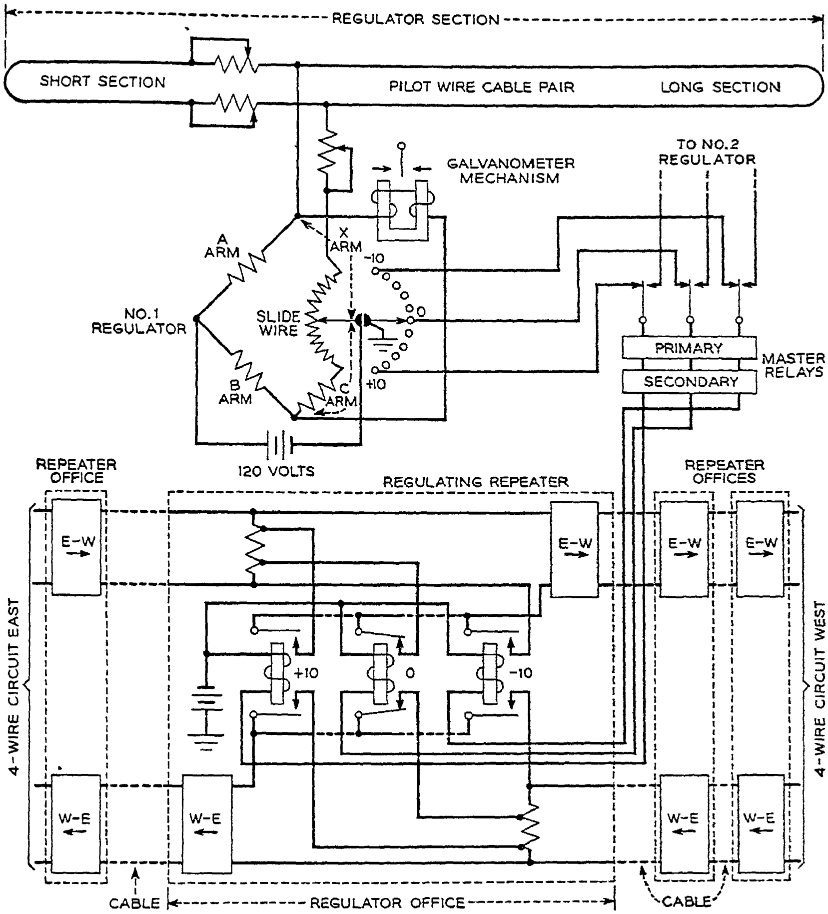


FIG. 8. Pilot Wire Transmission Regulator Circuit (Courtesy Bell System)

The desired overall transmission loss having been apportioned to the various component parts of toll connections, as shown under the plan, the pattern for the engineering of toll facilities is thus fixed. The *toll terminal loss* as determined for each toll center is defined as the average (one-half of the sum) of the effective transmitting and receiving losses (see article 15) from the toll circuit termination to (and including the efficiency of) the subscriber station apparatus. With this loss determined for a given toll center, the toll switching trunks and subscriber loops (exchange plant) must be engineered to meet this requirement, although exchange plant engineering also is subject to *exchange standards*.

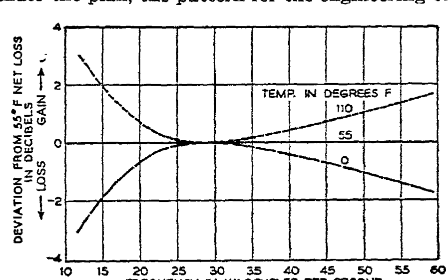


FIG. 9. Twist Effect in 100-mile Aerial Cable Circuit (Courtesy Bell System)

TOLL CIRCUIT LINE-UP PROCEDURE consists of adjusting the operating gains of voice frequency and carrier repeaters, carrier system terminals, and other associated apparatus, such as switching pads, equalizers, attenuators, and other devices necessary for proper operation of the toll circuits.

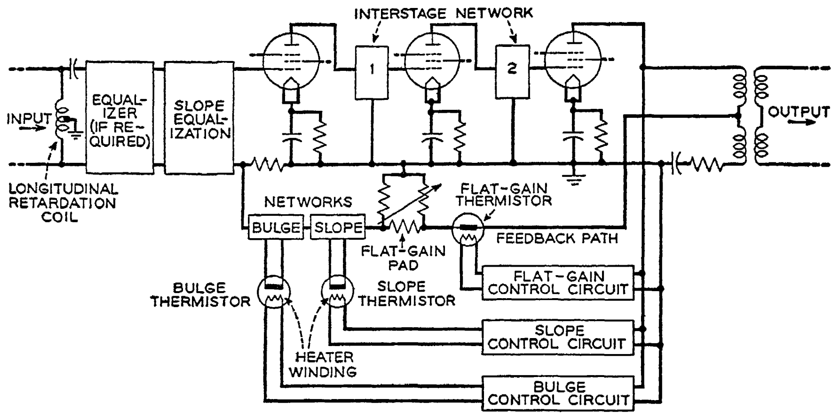


FIG. 10. Twist Correcting Circuit for Type K Carrier System (Courtesy Bell System)

From the general toll switching plan (Fig. 1) it is evident that intermediate toll circuits cannot be used for terminal traffic with the very low losses assigned without having excessive echo, singing, crosstalk, and other troubles. It is thus necessary to provide *switching loss pads* in each of these circuits at each end and in the end link circuits (PO-TC) at the primary outlet end, which pads remain in these circuits for terminal traffic and are automatically switched out of the circuits for via (through) traffic. The total pad loss usually employed is thus equal to the difference between the operating net loss of a given circuit in its terminal and via conditions.

In general, when a circuit meets the design objectives in the via condition and the loss of the switching pad or pads satisfies the crosstalk objectives in the terminal condition, singing will also be satisfactory; also, echo will be within limits if the pad loss is at least equal to the assigned negative echo margin.

Transmission levels on toll circuits require careful coordination with other adjacent or nearby circuits to prevent excessive crosstalk. Also levels should be relatively high (within the capabilities of the associated equipment) to provide suitable signal-to-noise ratios. Levels of about +3 to +6 db generally are employed at the input to two-wire lines, and from +4 to +10 db at the input to four-wire voice-frequency lines. Open-wire carrier system units employ about +16 db (input to the line) for Bell-owned systems.

TOLL LINE SIGNALING usually employs 20-, 135-, or 1000-cycle signaling systems. For short non-composited toll facilities, 20-cycle signaling is usual; for the longer toll circuits having composite sets 135- or 1000-cycle signaling is required. Since 1000-cycle signaling will readily pass along any type of message circuit that will transmit voice frequencies as such or in modulated form, this type of signaling is general for all long toll circuits. For crossbar toll operation, as described in article 3 of this section, pairs of frequencies in the range of 700 to 1700 cycles are sent out over toll trunks for pulsing the desired signals.

15. SERVICE REQUIREMENTS—EXCHANGE

EXCHANGE PLANT STANDARDS are based largely upon giving the telephone-using public a convenient, satisfactory service at the least cost consistent with protecting the investment and employee interests.

The establishing of *exchange plant standards* involves taking into consideration the efficiency of subscriber telephone sets as well as loop, trunk and central-office equipment (COE) losses and signaling ranges. It is thus necessary to establish a means of rating subscriber loops and trunks with respect to a known transmission standard in order that the capabilities and costs of the various types of equipment and line facilities that are available for use in exchange plant may be compared and it may be judged whether or in what respect the equipment facilities meet the assigned standards for a given exchange.

Overall transmission standards are the upper limits for the overall effective station-to-station transmission and form the basis of plant design. In the practical application of such standards, allowances are usually deducted for room noise, and sometimes for other impairments, from the overall standard so that the resulting *design standard* may be used directly in the design of the exchange plant.

The determination of the most desirable transmission standards for a given exchange is a matter of engineering and business judgment, based on a comprehensive view of local conditions and on past performance.

On the basis of general usage of the W.E. Co. FIA-AST or equivalent subscriber sets, Bell System practices contemplate ultimately overall exchange transmission standards of 10 to 14 db for multioffice exchange area traffic, with the possibility of the standards being 1 or 2 db higher for tandem operation. In single-office areas the standard is generally taken as approximately equal to the loop limit, considering both transmission and signaling.

A working reference system was devised some years ago by the Bell System as a means of rating exchange loop and trunk plant. This system, shown in Fig. 11, includes two identical common-battery subscriber loops, each connected through a 24-volt battery-feed repeating coil, to a variable, distortionless (up to 3000 cycles) 600-ohm impedance

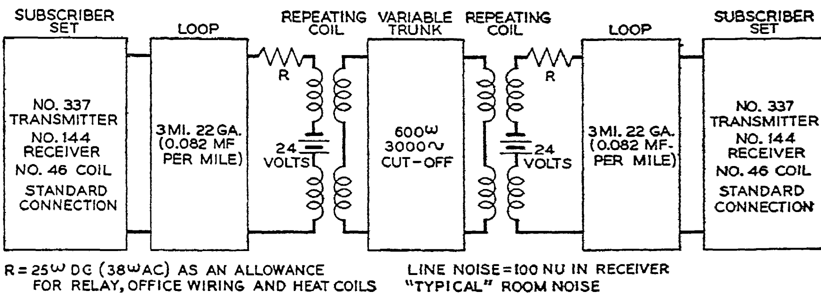


FIG. 11. Working Reference System for Specification of Effective Losses (Courtesy Bell System)

trunk. The length of loops and types of station and central-office apparatus are typical of conditions and apparatus existing at the time the system was devised and are still regarded as suitable for reference purposes.

The working reference system is so designed that it permits comparing *effective transmission losses*, which include volume and distortion losses, line and room noise, and sidetone effects. The system itself has an actual effective transmission loss or rating of 18 db, based on 7.5 db in the transmitting loop, 1.8 db in the receiving loop, and 8.7 db in the trunk. The line and room noises included in each loop are respectively 100 noise units (NU) and room noise comparable to that in quiet offices or fairly noisy residences. This room noise is equivalent to 50 db RAP (reference acoustic power, which is 10^{-16} watt of sound power per square centimeter at the listening ear).

EFFECTIVE TRANSMISSION PERFORMANCE of exchange telephone plant, as interpreted in the United States and some other countries, is evaluated by the generally accepted method of counting the number of requests to repeat words or sentences, in a short interval of time, made by talkers with average volume over the circuit to be evaluated. Other methods of evaluation, such as the "immediate appreciation" method, have been studied but have not been adopted in the United States.

For convenience, each effective loss is considered to have three components—volume, distortion, and sidetone losses. The distortion and sidetone losses in the working reference system *are considered to be zero for reference purposes*, and the effective loss of this system is thus numerically equal to the volume loss.

The individual effective losses which make up the effective loss of a complete circuit are:

1. Transmitting loop loss.
2. Receiving loop loss.
3. Trunk loss.
4. Terminal junction loss.
5. Central-office loss.
6. Intermediate junction loss.
7. Loss due to line noise.

The effective loss due to room noise is not considered a circuit loss, although it does affect conversations.

Losses 1 and 2 are determined by comparing the effect on conversation of the element or complete loop to be rated with the effect of the corresponding element or complete loop of the reference system, using the other components of the reference system to complete the talking circuit. Thus, since the effective loss of the transmitting loop of the reference system is 7.5 db, any transmitting loop which is substituted for the reference loop and which gives the same grade of service (repetition rate) also has an effective loss of 7.5 db. However, if the substituted loop gives the same grade of service after an increase of 2 db is made in the variable reference trunk loss, then the substituted loop has an effective loss 2 db lower than the reference loop, or if the reference trunk loss is adjusted to 2 db less than normal the substituted loop has an effective loss 2 db higher than the reference loop. Receiving loops are rated in the same manner, using the reference receiving loss of 1.8 db.

A given trunk may be rated by substituting it for the reference trunk and using the reference loops in the connection. The adjustment in the reference trunk loss determines the rating of the given trunk (called the *effective connecting circuit loss*) with respect to the reference trunk loss of 8.7 db, the fact that the assigned effective loss of the reference trunk is equal to its attenuation below 3000 cycles being kept in mind. The effective connecting circuit loss thus obtained is not in a useful form but may be made so by dividing it into an *effective trunk loss* and two *effective terminal junction losses* (one at each end), which latter are considered equal for this symmetrical circuit.

This division between trunk and junction losses is necessary for the practical establishment of a set of effective trunk loss curves, since for each type of trunk the terminal junction losses are different for each different combination of loops, sets, and central offices at the trunk terminals. The *effective trunk loss* consists of the volume attenuation of the trunk and that part of the distortion that is proportional to trunk length, thus permitting the establishing, for each type of trunk, of a value of effective trunk loss on a per mile basis. Each effective *terminal junction loss* includes a *volume reflection correction* plus one-half of that part of the *distortion loss* of the trunk that is not included in the effective trunk loss plus a correction for the effect on *sidetone* of the trunk impedance. The effective terminal junction losses can thus be considered as correcting factors which, when added to the sum of the other losses, will give the correct effective loss for a complete circuit.

The assumption that the effective connecting circuit loss (trunk rating) is made up of a constant plus a loss proportional to length is a good approximation for complete circuits containing a single type of trunk, except for effective trunk losses of about 5 db or less and for coil loaded trunks of any effective loss. For the low-loss trunks, accuracy in determining the trunk loss is not usually important. For loaded trunks, the curve of connecting circuit loss versus length departs from a straight line, because the end sections change with a change in the trunk length. However, if the connecting circuit loss is plotted only for lengths permitting half-section termination, a smooth curve is obtained which closely approximates a straight line above 5-db loss, and this curve may be used to determine the trunk loss per mile and the terminal junction loss for this termination. Similar curves, approximating straight lines, paralleling the one for half-section termination, may be set up for any other desired end section at each end. Such approximate straight lines determine the effective terminal junction losses for the respective end sections chosen. Thus, the *effective connecting circuit loss* of a loaded trunk, terminated at half-section at one end and at other than half-section at the other end, is considered as made up of the *effective trunk loss per mile* times the total geographical length of the trunk in miles (including the end sections) plus the *effective terminal junction loss* for each end section.

The effective loop losses apply directly only with the 600-ohm reference trunk. The effective trunk loss applies satisfactorily for any combination of loop, set, and central office, but the effective terminal junction losses obtained with the reference loops do not apply for other combinations of loop, set, and central office. The terminal junction losses for other combinations of these elements, including other than the reference trunk, may be determined as required.

EFFECTIVE LOOP LOSSES include the loss of three different types of *simplified central-office circuits*, namely, the 24-volt repeating coil circuit used in the working reference system, a 48-volt repeating coil circuit typical of standard circuits for toll connections, and a 48-volt step-by-step circuit used in local dial offices.

In practice, actual central-office circuits have equipment and wiring, additional to the simplified circuits, the loss of which is dependent upon the actual loop and trunk conditions. However, a single value of loss for the additional equipment and wiring for each type of office and connection is usually sufficient if it is determined under typical limiting loop conditions. The *effective local offices losses* are therefore determined between the working reference loops and trunk for the more commonly used central-office connecting circuits.

For trunks composed of different types of facilities, such as 19- and 22-gage cable, *intermediate junction losses* occur at each junction of the dissimilar types of facilities because of reflections of energy at these points. Such losses have been determined for various combinations of facilities used in the Bell System and must be added as part of the overall trunk loss.

Effective losses due to line noise (Table 3 of this section) should be added to the other effective losses in a given loop when the electrical noise at the receiver terminals of the loop is greater than the reference noise (100 units) assumed in the reference loops. Less noise than reference noise in the receiver is considered equivalent to reference noise.

In addition to the above-mentioned losses, if the room noise for a given loop is more or less than the room noise (50 db RAP) assumed for the reference loop, *effective room noise losses* corresponding to the difference between the actual and assumed room noise must also be added to the overall given loop loss. Figure 12 shows the effective loss due to room noise versus room noise of different intensities for two different sidetone conditions.

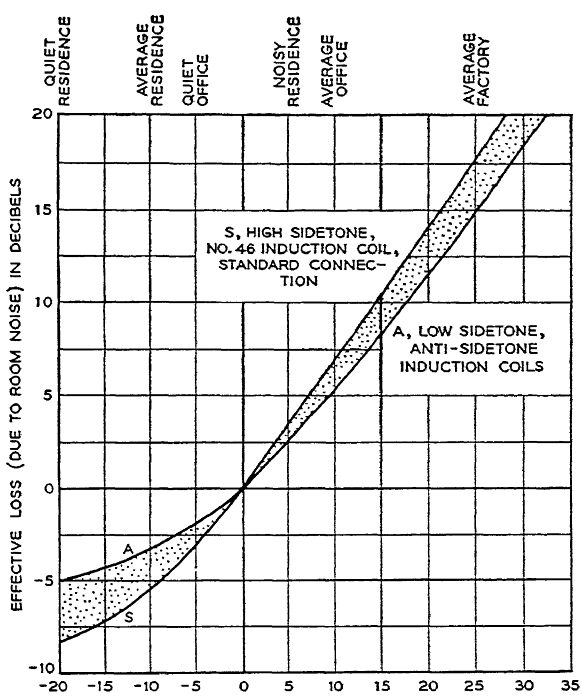


FIG. 12. Effective Receiving Loss Due to Room Noise, versus Room Noise of Different Intensities (Courtesy Bell System)

The values shown are approximate, since it is difficult to determine the actual losses due to room noise because of a number of variable factors, some of an intangible nature.

The Bell Telephone Laboratories have made field and laboratory investigations comparing the ease of carrying on a telephone conversation over different circuits in actual service, and also studying the physical characteristics of circuits and instruments, as well as their ability to transmit speech sounds, using syllabic articulation tests under a large number of variable conditions encountered in service. From these investigations a number of *effective transmission loss curves* showing effective loop losses in decibel versus loop length in thousands of feet have been prepared for the commonly used types of central offices, loop facilities (cable or open wire), and telephone sets, based on the working reference system.

Figure 13 shows separate effective transmitting and receiving loss curves (provisional) for both 24-volt exchange grade and 48-volt toll grade battery supply and also the average $(T + R)/2$ curves for both grades of battery supply. These particular curves apply for W.E. Co. type 1 or 10 Manual or Panel Dial Offices, non-loaded 22-BSA gage cable loops, and the latest-type W.E. Co. Antisidetone (AST) subscriber sets with FIA (handset) or 635-706A (deskstand) type instruments. These curves are typical of other sets of subscriber loop loss curves for different types of central offices, loop facilities, and sets, except that the loss values are different for the different conditions.

In addition to the non-loaded subscriber loop loss curves, effective loss curves are required showing, for different lengths and conditions, the losses of loaded trunks and loops, subscriber loop losses having loops composed of two or more different types of facilities, current supply losses versus transmitter current and versus loop resistance, transmitting and receiving losses due to sidetone, terminal junction losses, and many similar curves useful to the engineer.

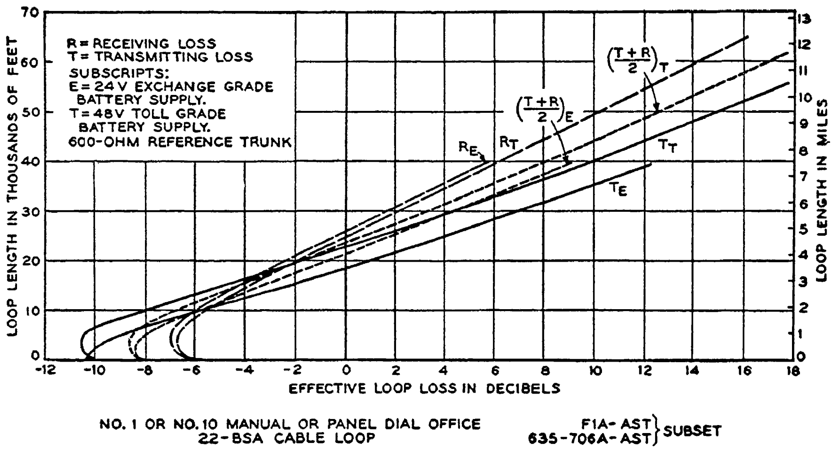


FIG. 13. Effective Transmission Losses in Common-battery Subloops (Courtesy Bell System)

SWITCHBOARD OPERATOR EFFECTIVE TRANSMISSION PERFORMANCE is also rated by means of the working reference system, except that the reference trunk impedance is 900 ohms instead of the 600 ohms employed for subscriber loop ratings. In making comparisons, the operator circuit is connected to the reference system in place of one of the subscriber loops. Since no loop is involved for this connection, single values of transmitting and receiving losses are adequate for rating each combination of operator's telephone set circuit and instruments.

The typical values of line and room noise specified for the operator terminal are 45 db RAP for line noise in the listening ear and 65 db RAP for room noise. The reference operator receiver (W.E. Co. No. 528) is more efficient than the reference subscriber receiver (W.E. Co. No. 144), and the reference room noise for operators is higher than that for subscribers. Incoming room noise from the distant operator's set also adds to the overall effective loss in the receiving operator's circuit. The latest design in operator transmitters and receivers provides improvements in operator transmission which will permit overall operator to operator standards in the approximate range of 10 to 12 db, which is comparable with present local subscriber to subscriber plant design.

16. PLANT DESIGN—TOLL

TOLL FACILITIES have been developed, since the earliest open-wire (iron) type, to include the following: (1) open wire, (2) cable, (3) carrier, and (4) radio.

TOLL OPEN-WIRE FACILITIES consist principally of hard-drawn copper, copper steel, and high-tensile-strength steel bare wire.

The three gages of *hard-drawn copper line wire* employed most commonly in Bell System plant are 165, 128, and 104 (mil diameter) wire, the electrical characteristics of which are shown for various positions and arrangements on the pole line in Fig. 14. For wet weather the value of g increases and other factors change. Several other gages of copper line wire are in use in this and other countries, such as 102 mil diameter, but the three mentioned above are representative of this type of wire usage. Also, a number of basic wire gages are in use, so that the mil diameter of a given gage of wire, such as No. 8, may be different under the different basic gages. Figures 15, 16, and 17 show, respectively, the attenuation frequency characteristics of different gages of hard-drawn copper physi-

Type of Circuit (hard-drawn copper)	(Gage of Wires, mils)	Spacing of Wires, in.	Constants per Loop Mile				Propagation Constant				Line Impedance				Wave-length, miles	Velocity, miles per second w	Trans-mission Loss, db/mi
							Polar		Rectangular		Polar		Rectangular				
			r, ohms	L, henrys	C, μf	g, m. mho	Magni-tude	Angle, degrees +	α	β	Magni-tude	Angle, degrees −	R, ohms	X, ohms −			
Non-pole pair side	165	12	4.11	0.00337	0.00915	0.29	0.0352	84.36	0.00346	0.0350	612	5.35	610	57	179.5	179,500	0.030
Pole pair side	165	18	4.11	.00364	.00863	.29	.0355	84.75	.00325	.0353	653	5.00	651	57	178.0	178,000	.028
Non-pole pair phantom	165	12	2.06	.00208	.01514	.58	.0355	85.34	.00288	.0354	373	4.30	372	28	177.5	177,500	.025
Pole pair phantom	165	18	2.06	.00207	.01563	.58	.0359	85.33	.00293	.0358	366	4.33	365	28	175.5	175,500	.025
Non-pole pair phys	165	8	4.11	.00311	.00996	.14	.0353	83.99	.00370	.0351	565	5.88	562	58	179.0	179,000	.032
Non-pole pair side	128	12	6.74	.00353	.00871	.29	.0356	81.39	.00533	.0352	650	8.32	643	94	178.5	178,500	.046
Pole pair side	128	18	6.74	.00380	.00825	.29	.0358	81.95	.00502	.0355	693	7.72	686	93	177.0	177,000	.044
Non-pole pair phantom	128	12	3.37	.00216	.01454	.58	.0357	82.84	.00445	.0355	401	6.73	398	47	177.0	177,000	.039
Pole pair phantom	128	18	3.37	.00215	.01501	.58	.0362	82.82	.00453	.0359	384	6.83	382	46	174.8	174,800	.039
Non-pole pair phys	128	8	6.74	.00327	.00944	.14	.0358	80.85	.00569	.0353	603	8.97	596	94	178.0	178,000	.049
Non-pole pair side	104	12	10.15	.00366	.00837	.29	.0363	77.93	.00760	.0355	692	11.75	677	141	177.0	177,000	.066
Pole pair side	104	18	10.15	.00393	.00797	.29	.0365	78.66	.00718	.0358	730	10.97	717	139	175.5	175,500	.062
Non-pole pair phantom	104	12	5.08	.00223	.01409	.58	.0363	79.84	.00640	.0357	421	9.70	415	71	176.0	176,000	.056
Pole pair phantom	104	18	5.08	.00222	.01454	.58	.0368	79.81	.00651	.0362	403	9.83	397	69	173.6	173,600	.056
Non-pole pair phys	104	8	10.15	.00340	.00905	.14	.0367	77.22	.00811	.0358	644	12.63	629	141	175.5	175,500	.070

Notes:
1. All values are for dry weather conditions.
2. All capacity values assume a line carrying 40 wires.
3. Resistance values are for temperature of 20 deg cent (68 deg fahr).
4. DP insulators assumed for all 12-in. and 18-in. spaced wires—CS insulators assumed for all 8-in. spaced wires.

FIG. 14. Characteristics of Open-wire Telephone Circuits at 1000 Cycles per Second. (Courtesy Bell System.)

cal (side in Fig. 15) circuits over the voice, type C, and type J carrier frequency ranges, for both wet and dry weather conditions. The wire spacing and types of insulators involved are 12-in. and double-petticoat (DP) for Fig. 15, and 8-in. and CS glass for Figs. 16 and 17. In the frequency range of 20 to 150 kc, the attenuation factor increases rapidly with frequency where the wires are covered with snow or ice. For example,

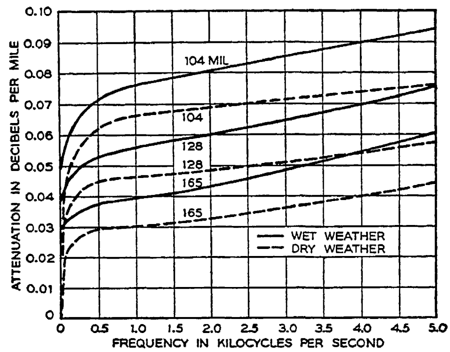


FIG. 15. Attenuation-frequency Characteristics of Open-wire Side Circuits over the Voice Range (Courtesy Bell System)

with about $1/2$ in. total diameter of melting glaze on a type J, 8 in., CS insulated, 165-gage carrier pair, the attenuation increases approximately from 0.13 db per mile at 20 kc to 0.9 db per mile at 150 kc. Variations of attenuation with temperature due to resistance change in open wire are about 1 per cent per 4 $1/2$ deg fahr change from 68 deg fahr.

In recent years, *copper steel wire* has been used extensively for telephone circuits, combining strength with relatively low transmission losses and d-c resistance values. This wire is manufactured with 30 and 40 per cent conductivity, which is the conductivity ratio (in percentage) of the wire to that of the annealed copper standard of like diameter.

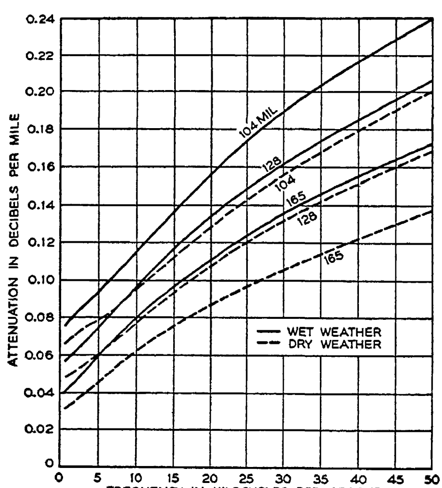


FIG. 16. Attenuation-frequency Characteristics of Open-wire Physical Circuits over the Type C Carrier Range (Courtesy Bell System)

A 40 per cent conductivity copper steel wire, which is the more commonly used type, has a steel core with a welded copper casing having a radial thickness of 20 per cent of the total radius of the wire.

The tensile strength and attenuation of copper steel pairs are about 2 to 2 $1/2$ times that of hard-drawn copper of the same size. Owing to higher attenuation, telephone repeaters

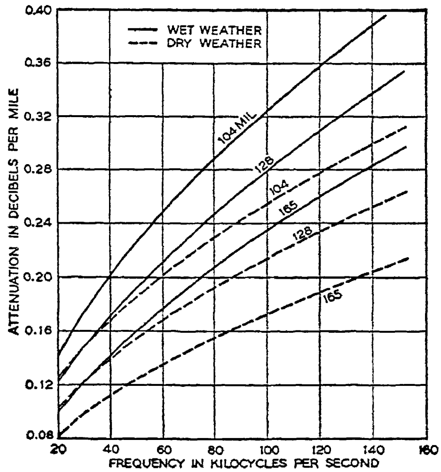


FIG. 17. Attenuation-frequency Characteristics of Open-wire Physical Circuits over the Type J Carrier Range (Courtesy Bell System)

require closer spacing with copper steel than with hard-drawn copper circuits. The characteristics of copper steel pairs are given in Table 5.

Table 5. Characteristics of Copper Steel Pairs

(Estimated for 68 deg Fahr—40 Per Cent Conductivity—53 Pairs CS Insulators per Mile)

Size of Wire, mil diameter	Frequency, kc	Resistance, ohms	Inductance, mh per pair mile		Characteristic Impedance, Dry, ohms		Attenuation, db per mile, dry		Attenuation, db per mile, wet	
			Pin Spacing		Pin Spacing		Pin Spacing		Pin Spacing	
			8 in.	12 in.	8 in.	12 in.	8 in.	12 in.	8 in.	12 in.
165	0.2	9.9	3.125	3.385	$761 - j516$	$793 - j542$	0.057	0.054	0.063	0.060
	1.0	10.3	3.060	3.320	$572 - j143$	$615 - j143$	0.078	0.073	0.084	0.078
	3.0	10.8	3.027	3.287	$558 - j\,51$	$597 - j\,51$	0.085	0.079	0.092	0.086
	10.0	11.4	2.995	3.255	$552 - j\,16$	$592 - j\,16$	0.092	0.085	0.104	0.098
	30.0	12.6	2.986	3.246	$551 - j\,6$	$591 - j\,6$	0.103	0.095	0.127	0.120
	140.0	24.7	2.974	3.234	$550 - j\,2.4$	$590 - j\,2.4$	0.207	0.192	0.282	0.273
128	0.2	16.6	3.285	3.545	$943 - j736$	$991 - j759$	0.077	0.073	0.084	0.081
	1.0	17.3	3.250	3.510	$635 - j230$	$674 - j232$	0.120	0.112	0.126	0.119
	3.0	18.0	3.206	3.466	$593 - j\,86$	$634 - j\,86$	0.134	0.124	0.142	0.132
	10.0	18.7	3.167	3.427	$583 - j\,27$	$625 - j\,27$	0.142	0.131	0.155	0.145
	30.0	19.7	3.148	3.408	$580 - j\,9.4$	$623 - j\,9.4$	0.152	0.139	0.177	0.168
	140.0	30.5	3.139	3.399	$580 - j\,3$	$622 - j\,3$	0.241	0.224	0.321	0.310
104	0.2	25.0	3.430	3.690	$1139 - j958$	$1189 - j988$	0.096	0.091	0.105	0.101
	1.0	25.7	3.410	3.670	$691 - j324$	$736 - j328$	0.162	0.152	0.169	0.160
	3.0	26.6	3.357	3.617	$621 - j124$	$666 - j124$	0.188	0.174	0.196	0.183
	10.0	27.7	3.313	3.573	$606 - j\,40$	$651 - j\,40$	0.201	0.186	0.215	0.200
	30.0	28.7	3.287	3.547	$602 - j\,14$	$647 - j\,14$	0.211	0.195	0.238	0.224
	140.0	37.6	3.277	3.537	$602 - j\,3.8$	$647 - j\,3.8$	0.284	0.264	0.367	0.353

Notes:

1. Resistance (d-c) is 0.1 ohm less than resistance at 0.2 kc for all three gages.
2. Leakage conductance and capacities of copper steel pairs are comparable to those of same size hard-drawn copper pairs (Fig. 14).
3. For DP insulators the attenuation change from dry to wet weather conditions is about twice that for CS insulators.
4. The above estimated values will vary somewhat from actual measurements, owing to the effects of transpositions, spacing, and other small irregularities, which cannot be calculated. For 12-in. pairs, assume on the average about 10 per cent lower high-frequency impedance and about 10 per cent higher high-frequency attenuation values. The deviations should be less for the closer spaced pairs with point-type transpositions. (Courtesy Bell System.)

HIGH-STRENGTH STEEL WIRE has practically replaced the various grades of iron (E.B.B. and B.B.) and mild steel wire for telephone purposes because of its greater strength and coordination with other wire services. The high-strength steel wire now in production by several manufacturers is used for only the very short toll circuits in some parts of the country, on account of its high attenuation and inherent noise characteristics. Its principal usage is in exchange plant. Steel wire is now generally zinc coated (galvanized) electrolytically, which insures a more uniform coating than the "hot dip" process. The usual weights of zinc coating vary from 0.8 to 2.4 oz per sq ft, depending upon the customer's requirements. The characteristics are given in more detail in article 17, Plant Design—Exchange.

CABLE FACILITIES for toll purposes comprise a number of types of loaded and non-loaded cable pairs. *Toll entrance cables* are employed for extending open-wire toll lines into toll or toll terminal offices or stations for distances usually limited to a few miles. *Toll cables* are employed as permanent backbone toll plant, interconnecting principal cities and important intermediate switching and equipment centers. Toll and toll entrance cables may be of the aerial or underground type or combinations of both, but the tendency is to place them underground for greater reliability of service.

Table 6. Load-coil Spacings

Code Designation	Nominal Spacing, feet	General Application
A	700 *	Cables serving open-wire carrier.
B	3000 *	Cables serving open-wire carrier. Toll and exchange cables.
C	929 *	Cables serving open-wire carrier.
D	4500 *	Exchange cables.
E	5575 *	Toll entrance cables. Replaced by H, except when used with C spacing.
F	2787 *	Cables serving open-wire carrier (phantom circuits).
H	6000 *	Toll, toll entrance, and exchange cables.
J	600 †	Cables serving J open-wire carrier.
X	680 *	Equivalent capacity in carrier office cables.
Y	2130 *	Equivalent capacity in carrier office cables.

* For side and phantom cable capacitances of 0.062 and 0.102 μf per mile, respectively.

† For physical pair capacitance of 0.025 μf per mile.

Cable facilities, being necessarily of small-gage, soft-drawn, insulated copper conductors, in order to provide economical sizes of conductor complements, have relatively high mutual capacitances, resistances, and greatly increased attenuation per mile over the larger-gage, wider-spaced, open-wire conductors. Figure 18 shows the characteristics of standard types of paper-insulated cable telephone circuits at 1000 cycles per second, as used in the Bell System. In column 2 of Fig. 18, the abbreviations N.L.S. and N.L.P. refer to non-loaded side and non-loaded phantom respectively, while the remaining abbreviations show the load coil spacing by letter (H-6000 feet and B-3000 ft), the inductance in millihenries by the figure, and whether the circuit is a side or phantom by the letters S and P respectively. For a designation such as H-44-25, the first and second figures refer to the side and phantom coil inductances, respectively.

Table 7. Toll-entrance Cable Loading

Carrier Frequency Range, kc	Type of Loading		
0– 10	B H-15-15 (B for side, H for phantom)		
0– 30	C E-4.1-12.8 (C for side, E for phantom and 165 open-wire circuits, 12 in.)		
0– 30	C E-4.8-12.8 (C for side, E for phantom and 128 or 104 open-wire circuits, 12 in.)		
0–145	J-0.72	165 open-wire, 8 in. 128 open-wire, 8 in.	Used with disk-insulated pairs in shielded, spiral-four quadded cable.
0–145	J-0.85	128 open-wire, 8 in.	
0–145	J-0.94	104 open-wire, 8 in.	
0– 30	X-2.7	Office cable loading.	
0– 10	Y-9	Office cable loading.	

The type of *loading* at present employed in *toll cables* for *four-wire* voice frequency message cable circuits is usually the 19-gage H-44-25 or H-88-50 type, since less variation in attenuation over the voice range and higher cutoff frequencies are obtained with these loadings (Fig. 18) than are possible with the older, higher-inductance-type loading. Repeaters generally are spaced at 40–50 mile intervals. For *two-wire* voice-frequency message cable circuits, 19-gage H-88-50 or B-88-50 loading is usually employed. The variation in attenuation over the voice range is small, the cutoff frequency is amply high, and telephone repeater spacings of about 40–50 miles generally are used.

For program transmission (article 18) 16-gage B-22 loaded cable pairs were specially

Wire Gage A.W.G.	Type of Loading	Spacing of Load Coils, miles	Load Coil Constants per Load Section: r, ohms	Load Coil Constants per Load Section: L, henrys	Constants Assumed to be Distributed per Loop Mile: r, ohms	L, henrys	C, μf	g, mmho	Propagation Constant, Rectangular: α	β	Line Impedance, Polar: Magnitude	Angle, degrees −	Wave-length, miles	Velocity, miles per second w	Cut-off Frequency f_c (approx.)	Transmission Loss db/mi calc.
19	N.L.S.	......	......	......	85.8	0.001	0.062	1.5	0.1249	0.1338	470.1	42.80	46.93	46,930	....	1.08
19	H-31-S	1.135	2.7	0.031	88.2	0.028	.062	1.5	.0643	.2693	710.0	13.20	23.33	23,331	6,700	0.56
19	H-44-S	1.135	4.1	.043	89.4	.039	.062	1.5	.0561	.3138	818.0	9.91	20.02	20,022	5,705	.49
19	H-88-S	1.135	7.3	.088	92.2	.078	.062	1.5	.0418	.4388	1,131.0	5.22	14.32	14,319	3,997	.36
19	H-172-S	1.135	13.0	.170	97.3	.151	.062	1.5	.0323	.6085	1,564.7	2.82	10.33	10,326	2,878	.28
19	H-174-S	1.135	16.1	.171	100.0	.152	.062	1.5	.0331	.6107	1,570.0	2.84	10.29	10,288	2,870	.29
19	H-245-S	1.135	24.5	.247	107.4	.219	.062	1.5	.0300	.7236	1,882.0	2.12	8.58	8,577	2,389	.26
19	B-88-S	0.568	7.3	.088	98.7	.156	.062	1.5	.0322	.6186	1,590.2	2.76	10.16	10,157	5,655	.28
16	N.L.S.	......	......	......	42.1	.001	.062	1.5	.0842	.0974	330.7	40.65	64.51	64,506	....	.73
16	H-31-S	1.135	2.7	.031	44.5	.028	.062	1.5	.0334	.2638	682.5	6.99	23.82	23,818	6,700	.29
16	H-44-S	1.135	4.1	.043	45.7	.039	.062	1.5	.0296	.3134	808.0	5.17	20.05	20,048	5,705	.26
16	H-88-S	1.135	7.3	.088	48.5	.078	.062	1.5	.0224	.4374	1,124.0	2.71	14.36	14,365	3,997	.19
16	H-172-S	1.135	13.0	.170	53.6	.151	.062	1.5	.0183	.6082	1,562.0	1.51	10.33	10,331	2,878	.16
16	H-174-S	1.135	16.1	.171	56.3	.152	.062	1.5	.0191	.6102	1,567.0	1.58	10.30	10,297	2,870	.17
16	H-245-S	1.135	24.5	.247	63.7	.219	.062	1.5	.0184	.7323	1,880.0	1.22	8.58	8,580	2,389	.16
16	B-88-S	0.568	7.3	.088	54.9	.156	.062	1.5	.0185	.6181	1,587.4	1.49	10.17	10,165	5,655	.16
19	N.L.P.	......	......	......	42.9	.0007	.100	2.4	.1106	.1219	262.1	41.97	51.53	51,525	....	.96
19	H-18-P	1.135	1.4	.018	44.1	.017	.100	2.4	.0529	.2642	428.8	11.11	23.78	23,781	6,959	.46
19	H-25-P	1.135	2.1	.025	44.7	.023	.100	2.4	.0466	.3047	490.7	8.48	20.62	20,621	5,916	.40
19	H-50-P	1.135	3.7	.050	46.2	.045	.100	2.4	.0351	.4228	675.2	4.53	14.86	14,861	4,193	.30
19	H-63-P	1.135	6.1	.063	48.3	.056	.100	2.4	.0331	.4712	751.8	3.80	13.33	13,334	3,738	.29
19	H-106-P	1.135	8.2	.107	50.1	.095	.100	2.4	.0269	.6129	976.4	2.29	10.25	10,252	2,871	.23
19	H-155-P	1.135	12.5	.155	53.9	.137	.100	2.4	.0244	.7357	1,171.6	1.68	8.54	8,540	2,386	.21
19	B-50-P	0.568	3.7	.050	49.4	.089	.100	2.4	.0273	.5933	945.2	2.41	10.59	10,590	5,936	.24
16	N.L.P.	......	......	......	21.0	.0007	.100	2.4	.0746	.0890	184.8	38.98	70.60	70,604	....	.65
16	H-18-P	1.135	1.4	.018	22.2	.017	.100	2.4	.0273	.2604	416.7	5.76	24.13	24,129	6,959	.24
16	H-25-P	1.135	2.1	.025	22.8	.023	.100	2.4	.0243	.3022	482.5	4.37	20.79	20,792	5,916	.21
16	H-50-P	1.135	3.7	.050	24.3	.045	.100	2.4	.0189	.4218	672.1	2.35	14.90	14,896	4,193	.16
16	H-63-P	1.135	6.1	.063	26.4	.056	.100	2.4	.0185	.4705	749.4	2.04	13.35	13,354	3,738	.16
16	H-106-P	1.135	8.2	.107	28.2	.095	.100	2.4	.0156	.6126	975.2	1.24	10.26	10,257	2,871	.14
16	H-155-P	1.135	12.5	.155	32.0	.137	.100	2.4	.0151	.7355	1,170.9	0.95	8.54	8,543	2,386	.13
16	B-50-P	0.568	3.7	.050	27.5	.089	.100	2.4	.0157	.5929	943.9	1.30	10.60	10,597	5,936	.14
16	B-22	0.568	1.25	.022	43.1	.040	.062	1.5	.0273	.3139	809.1	4.76	20.01	20,010	11,276	.24

Note. The values for cutoff frequency and transmission loss per mile, as given in the last two columns, are calculated from the primary constants which are assumed as uniformly distributed. These values accordingly may not be identical with the measured values given in standard formal instructions. The values given in *Bell System Practices* should therefore be used for engineering work.

FIG. 18. Characteristics of Standard Types of Paper-insulated Cable Telephone Circuits at 1000 Cycles per Second (Courtesy Bell System)

developed to give a high cutoff frequency. These facilities are capable of transmitting a frequency band up to about 8000 cycles without serious distortion.

Figure 19 shows the attenuation frequency characteristics of various types of cable circuits loaded for voice frequency and program service.

Table 6 gives the code designations of load coil spacings as devised for Bell System use.

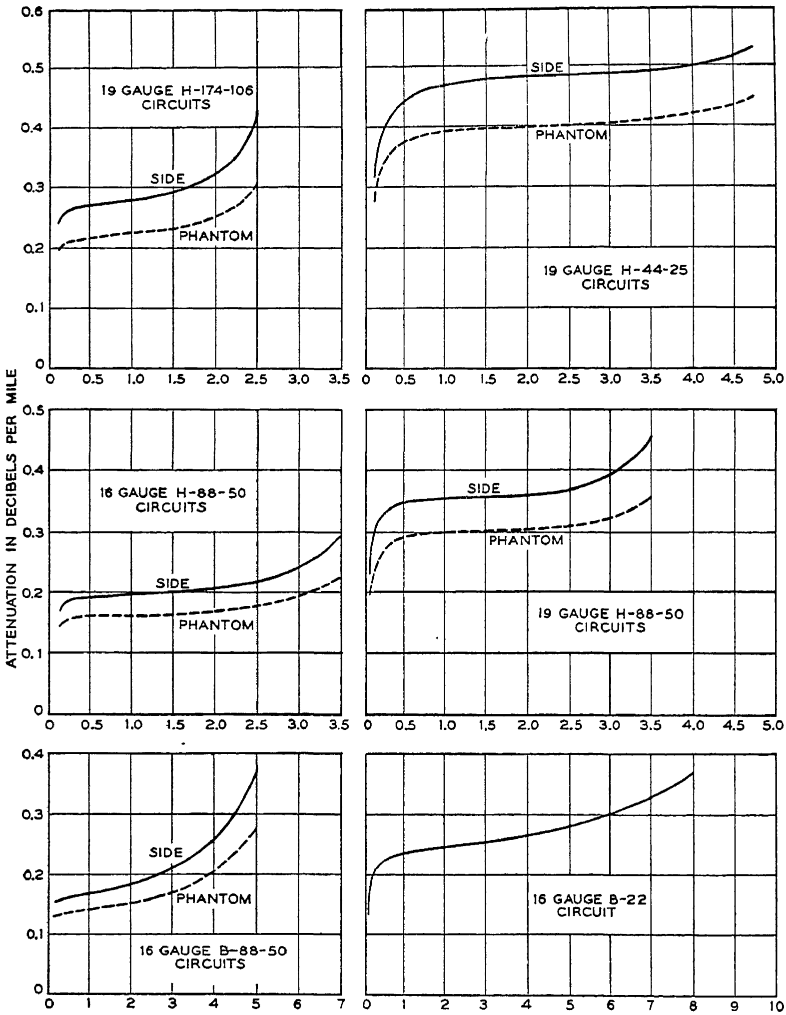


FIG. 19. Attenuation-frequency Characteristics of Various Types of Loaded Cable Circuits (Courtesy Bell System)

The type of *loading* generally used in *toll entrance cables* is of the H-31-18 type for voice-frequency circuits and of the types shown in Table 7 for carrier circuits.

Non-loaded cable facilities are employed in both toll entrance and toll cables. Such toll entrance facilities may be used without appreciable transmission penalty up to about 3000 ft in extending open-wire circuits into offices or intermediate toll equipment points. For K carrier systems, which are operated through toll cables, only non-loaded pairs may be used, since loading is not available for the K frequency range. For J carrier systems a specially designed low-capacity, 16-gage, spiral-four loaded or non-loaded cable may be used for entrances.

Cable facilities used for toll purposes have, in general, been composed of 10-, 13-, 16-, and 19-gage cable conductors, paired and quadded for physical and phantom circuit operation. Some non-quadded cables for physical circuits only have been used. At present 19 gage is largely employed, but in a few cases 16 gage may be added. The newer *toll cables* have the short pair twist, high dielectric strength, core to sheath, and nominal mutual capacities of 0.062 and 0.096 μf per mile for side and phantom respectively. The standard sizes of toll and toll entrance cable cover a wide range of standard numbers of quads and pairs, and several different gages may be placed in the same lead cable sheath, including pairs for exchange use, as required to meet service needs.

Toll entrance cables for carrier, non-phantomed, open-wire lines have the long pair twist and nominal capacities of 0.062 and 0.102 μf per mile for the side and phantom, respectively.

Variations in net losses in cable facilities due to temperature changes are shown in Fig. 20.

Loading	Loss at 55 Deg Fahr, decibels per mile				
	22-gage	19-gage	16-gage	13- or 14-gage *	10-gage
B-22-N	....	0.45	0.236	0.157	....
B-44-N	....	.34	....	....	....
B-88-50-S	0.50	.28	.16	....	....
B-88-50-P	.42	.23	.14	....	....
H-22-N	....	.62	.32	....	....
H-44-25-S	.92	.47	.25	.14	0.08
H-44-25-P	.77	.39	.21	.115	.07
H-88-50-S	.66	.35	.19	.12 *	....
H-88-50-P	.56	.30	.16	....	....
H-172-63-S, H-174-63-S	.49	.27	.16	....	....
H-172-63-P, H-174-63-P	.51	.28	.16	....	....
H-174-106-S	.50	.28	.16	.10	....
H-174-106-P	.40	.22	.13	.084	....

* Value marked with * applies to 14-gage.

Loading	± Variations from Loss at 55 Deg Fahr, decibels per mile					
	22-gage		19-gage		16-gage	
	Aerial Cable *	U.G. Cables †	Aerial Cable *	U.G. Cable †	Aerial Cable *	U.G. Cable †
B-22-N	....	....	0.052	0.017	0.028	0.009
B-44-N	....	....	.040	.013	....	....
B-88-50-S	0.060	0.020	.031	.011	.018	.006
B-88-50-P	.050	.017	.026	.009	.015	.005
H-22-N	....	....	.071	.024	.037	.012
H-44-25-S	.110	.037	.055	.018	.029	.010
H-44-25-P	.092	.031	.046	.015	.024	.008
H-88-50-S	.079	.026	.041	.014	.022	.007
H-88-50-P	.067	.022	.035	.012	.019	.006
H-172-63-S, H-174-63-S	.059	.020	.031	.010	.018	.006
H-172-63-P, H-174-63-P	.061	.020	.032	.011	.018	.006
H-174-106-S	.060	.020	.032	.011	.017	.006
H-174-106-P	.048	.016	.025	.008	.013	.004

* Temperature range, ±54 deg fahr; resistance variation, ±12 per cent
† Temperature range, ±18 deg fahr; resistance variation, ±4 per cent

FIG. 20. Net Loss Variations with Temperature for Different Gages and Loadings of Cable Circuits (Courtesy Bell System)

The cables may have a plain lead sheath covering with no outer protection, or, if protection is required from gophers, digging, or other extraneous disturbances, the sheath may be covered with jute, gopher and jute protection, corrosion protection, or tape armor for either aerial or buried construction. Submarine cables have single or double armored protection, composed of heavy steel wires laid around the sheath with jute and compound in one or two layers. Metallic shields are also used, as required, around the cable core, individual groups or quads of cable conductors, to protect circuits from electrical disturbances.

CARRIER FACILITIES may be provided in either open wire or cable. For any given carrier system between two points, the facilities selected must be suitable for that system. *Low-frequency carrier systems* of the G and H types generally employ 12-in.-spaced, open-wire facilities, having voice-frequency characteristics, except that, as the number of these systems increase on a given pole line, a carrier transposition scheme may be required to limit crosstalk, since the usual voice-frequency schemes are not designed for the higher frequencies. The C carrier system generally employs 12- or 8-in.-spaced, open-wire facilities, transposed for frequencies up to 30 kc, although the physical pairs or phantom groups may be transposed one at a time, according to the selected transposition scheme, as required to provide for the carrier systems on the given pole line. The J carrier system, being in the *high-frequency group*, requires 8- or 6-in.-spaced, open-wire facilities specially transposed to limit crosstalk at frequencies up to 140 kc.

With the G, H, and C systems, DP insulators and transposition drop brackets are usually satisfactory for wet or dry weather, but for the J system it is necessary to provide CS insulators and point-type transpositions, where the pole line will initially or ultimately have a full complement of carrier systems. Wire sag must be limited to small deviations from normal and all other physical irregularities must be controlled where carrier systems are operated. These limitations assume greater importance as the system frequency increases.

The K carrier system, operating through cable only, requires suitable conductors. Non-loaded pairs of 0.062-μf capacity per mile and of 19 gage, selected as the most economical gage, must be properly balanced in the cable mutually and against other facilities and segregated (using separate cables or layer shields) for the two directions of transmission. Figure 21 shows the attenuation-frequency characteristics of 19-gage non-loaded cable circuits over the K carrier range at different temperatures.

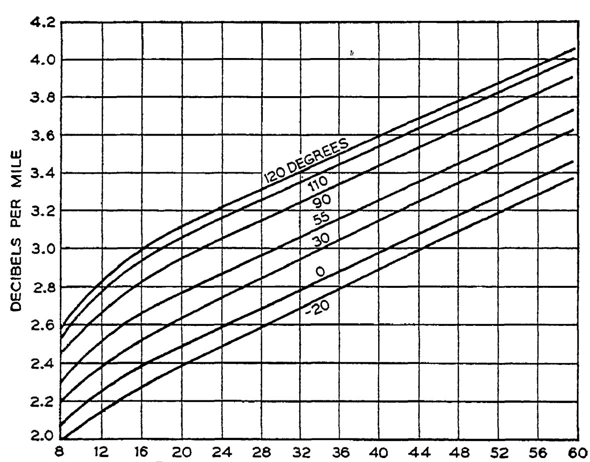


FIG. 21. Attenuation-frequency Characteristics of Non-loaded 19-gage Cable Circuits over the Type K Carrier Range at Different Temperatures (Deg Fahr) (Courtesy Bell System)

COAXIAL CABLE is required for L carrier system operation, on account of the high-frequency range of this system (up to possibly 7 Mc or more). This type of cable, first installed commercially between Minneapolis, Minn., and Stevens Point, Wis., in 1941, is built up of multiples of two coaxial tubes or units, plus ordinary paper-insulated conductors as needed to take care of requirements for short-haul message circuits and signaling and alarm trunks.

The Stevens Point-Minneapolis cable contained four coaxial units, with two 22-gage pairs in the center, a 19- and 22-gage pair in each of the four outer interstices between the units, and eighteen 19-gage quads surrounding this assembly. Since that installation many changes and improvements have been made in the design and construction of coaxial cable. One of the latest designs provides for 8 coaxial units and about 78 quads of 19-gage copper equivalent, built in with and surrounding the units to form a full-size lead-sheathed cable. The quads may be used for message and program service, order wires, alarms, and miscellaneous applications, and the coaxial units for message, television, and possibly other services.

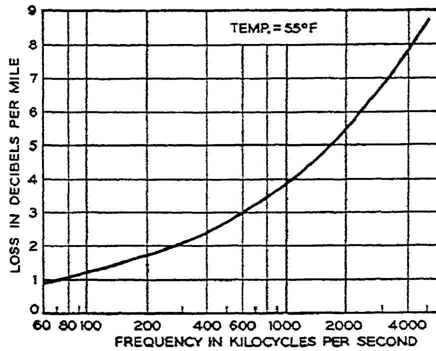


Fig. 22. Attenuation-frequency Characteristics of an 0.375-in. Longitudinal Seam Coaxial Cable Circuit (Courtesy Bell System)

The present *coaxial unit* consists of a single 12-mil copper tape (outer conductor) formed into a tubular conductor, which is, in turn, tightly wrapped with two 6-mil steel tapes, the outer of which covers the gaps between the turns of the inner steel tape. Each edge of the copper tape has small serrations which engage the opposite edge of the copper tape, and the two edges thus interlock to form a tube. This design is known as "longitudinal seam" coaxial. A 10-gage (100.4-mil) copper wire (inner conductor) is supported in the center of this tube by polyethylene disks (0.085 in. thick), spaced about 1 in. apart.

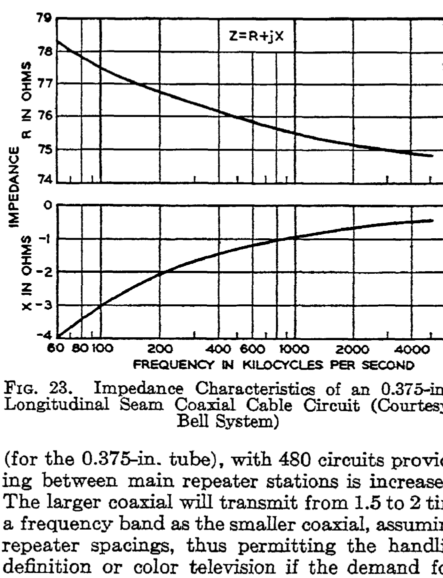


Fig. 23. Impedance Characteristics of an 0.375-in. Longitudinal Seam Coaxial Cable Circuit (Courtesy Bell System)

The inside diameter of the present standard coaxial tube is 0.375 in. (an original design employed a 13-gage central conductor and an outer conductor having an inside diameter of 0.27-in.). Attenuation and impedance characteristics are shown in Figs. 22 and 23. The new type permits lengthening auxiliary repeater spacings from about 5.4 miles (for the 0.27-in. tube) to about 7.9 miles (for the 0.375-in. tube), with 480 circuits provided in each case. Also the maximum spacing between main repeater stations is increased from about 90 to more than 150 miles. The larger coaxial will transmit from 1.5 to 2 times as wide a frequency band as the smaller coaxial, assuming the same repeater spacings, thus permitting the handling of high definition or color television if the demand for it arises. Figure 24 shows construction of an eight-unit coaxial with a complement of wire quads.

Lightning protection for the complete cable is provided by placing a 10-mil, corrugated, copper jacket over the lead sheath of the cable, which is first wrapped in thermoplastic material and a layer of tough cloth. The copper jacket is then covered with cloth tape.

Each regular coaxial unit is paired with an identical alternate coaxial unit, which is automatically switched into service, replacing the regular unit, in case of fault occurrence.

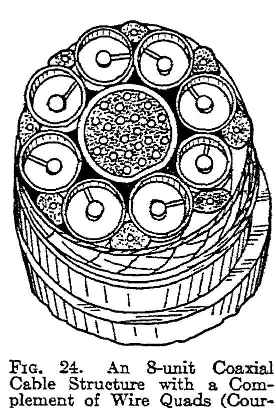
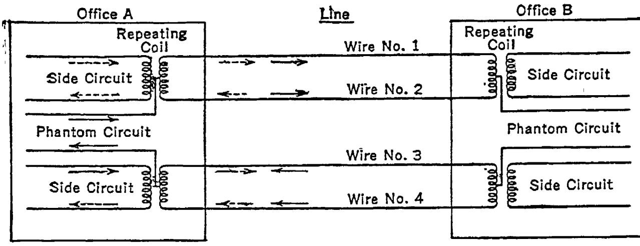


Fig. 24. An 8-unit Coaxial Cable Structure with a Complement of Wire Quads (Courtesy Bell System)

PHANTOM CIRCUIT operation is common in both open wire and cable. Two open-wire physical circuits or two cable pairs, which are twisted together to form a quad (four wires), can be equipped at their terminals to provide a third circuit, called a phantom, as shown in Fig. 25.

The side-circuit currents circulate over the side circuits as shown by the dashed arrows, and the phantom-circuit currents circulate over the phantom circuit as shown by the solid arrows. The line sides of the phantom repeating coils are closely balanced with respect to the phantom tap so that the phantom-circuit current, dividing about equally



←– – – instantaneous direction of side-circuit currents.
←——— instantaneous direction of phantom-circuit currents.

FIG. 25. Phantom Circuit Derived from Two Side Circuits

between the two wires of each pair, causes no appreciable induced current in the side-circuit terminations. Also, the side-circuit currents do not enter the phantom. This arrangement provides three circuits from four wires but requires a well-maintained balance in the line sides of the repeating coils and between the four wires, and adequate transposing, both in the open-wire group and the cable quad, to prevent excessive noise and crosstalk.

17. PLANT DESIGN—EXCHANGE

Exchange facilities, employed in connecting subscriber station equipment with central offices, for interoffice trunking in multioffice exchange areas, and for miscellaneous uses, may consist of cable or open wire or a combination of both types of facilities. The latest

Gage Cable		26		24			22				19		16	13
Type of cable		ST AST	BST	M SM ASM CSM	DSM	NM	SA ASA BSA CSA	NA ANA	TA	TS	BNB CNB	TB ANB DNB	TH NH	TJ
R, ohms per mile at 68° F		440	440	274	274	274	171	171	171	171	85	85	42	21.4
C, μf per mile		.069	.079	.072	.084	.065	.082	.073	.062	.068	.084	.066	.066	.066
Loading	Load Spacing, feet	Decibels per Mile at 68 Deg Fahr												
B-175	3,000	.94	1.01	.63	.68	......	.44	......	......	......	.25	.22	.14	......
B-135	3,000	1.05	1.12	.69	.75	......	.48	......	......	......	.26	.24	.14	......
B-88	3,000	1.30	1.39	.87	.94	......	.60	......	......	......	.34	.30	.18	......
D-175	4,500	1.12	1.20	.74	.80	......	.51	.49	......	......	.28	.25	.15	......
D-135	4,500	1.25	1.33	.82	.88	......	.56	......	......	......	.30	.27	......	......
D-88	4,500	1.52	1.62	1.01	1.09	......	.70	......	......	......	.38	.34	......	......
H-250	6,000	......	......	......	......	......	......	......	......	......	......	.26	.17	.11
H-175	6,000	......	......	......	......	......	......	......	......	......	.31	.27	.15	.10
H-135	6,000	1.40	1.50	.92	1.00	......	.63	.60	......	......	.34	.30	.16	......
H-88	6,000	1.69	1.80	1.14	1.23	......	.79	......	......	......	.42	.38	.21	......
H-44	6,000	2.06	2.21	1.46	1.58	......	1.04	......	......	......	.56	.50	.27	......
M-175	9,000	......	......	......	......	......	......	.65	.60	.63	......	.33	.17	.11
M-135	9,000	1.63	1.75	1.09	1.18	1.14	.75	.73	......	......	.41	.36	.20	......
M-88	9,000	1.91	2.04	1.31	1.42	1.25	.92	.87	......	......	.49	.44	.24	.14
R-133	11,600	......	......	......	......	......	......	......	.76	.80	......	.41	.21	.12
Non-loaded	......	2.67	2.86	2.14	2.31	2.04	1.79	1.69	1.55	1.63	1.26	1.11	.75	.50

FIG. 26. Attenuation Losses of Non-quadded Exchange Area Cable Facilities

developments affecting rural service include (1) trial installations of telephone carrier systems on rural power distribution lines, serving areas where telephone facilities are not available, and (2) direct radio channels between the central office and farm or ranch homes heretofore economically inaccessible for the usual pole line and open-wire construction.

EXCHANGE CABLES, in general, employ five different gages of soft-drawn copper conductors, namely, 26, 24, 22, 19, and 16 gage. A relatively small amount of 28-gage cable, developed to conserve copper during World War II, has been manufactured, but with normal copper prices its use in place of 26-gage, where applicable, does not effect appreciable cost savings. Exchange trunk cables are usually of 19 or smaller gage.

Type of Facility	Values Shown Are: Per Unit Length of	At Temp of F°	Dry or Wet	R (Loop), ohms (dc)	L, henrys	G, mhos (1000 cycles) ($\times 10^{-6}$)	C, farads ($\times 10^{-6}$)	$\frac{G}{C}$
Non-quadded exchange area cables								
26 gage ST AST	Mile	68	......	440	0.001	1.8	0.069	26
26 gage BST	"	68	......	440	.001	2.1	.079	26
24 gage M SM ASM CSM	"	68	......	274	.001	1.9	.072	26
24 gage DSM	"	68	......	274	.001	2.2	.084	26
24 gage NM	"	68	......	274	.001	1.7	.065	26
22 gage SA ASA BSA CSA	"	68	......	171	.001	2.1	.082	26
22 gage NA ANA	"	68	......	171	.001	1.9	.073	26
22 gage TA	"	68	......	171	.001	1.6	.062	26
22 gage TS	"	68	......	171	.001	1.7	.068	26
19 gage BNB CNB	"	68	......	85	.001	2.2	.084	26
19 gage TB ANB DNB	"	68	......	85	.001	1.7	.066	26
16 gage TH NH	"	68	......	42	.001	1.7	.066	26
13 gage TJ	"	68	......	21.4	.001	1.7	.066	26
Submarine cables—non-quadded								
Single paper insulation 24 gage	Mile	55	......	266	.001	1.7	.066	26
Single paper insulation 22 "	"	55	......	166	.001	1.9	.075	26
Single paper insulation 19 "	"	55	......	83	.001	2.0	.078	26
Double paper insulation 24 gage	"	55	......	266	.001	1.8	.071	26
Double paper insulation 22 "	"	55	......	166	.001	2.1	.080	26
Double paper insulation 19 "	"	55	......	83	.001	2.2	.083	26
Double paper insulation 16 "	"	55	......	41	.001	1.7	.066	26
17-gage U Wire								
U bridle wire	Kilofoot	68	Wet	10.3	.00033	*	.025	......
U / UA distribution wire (buried)	Kilofoot	68	"	10.3	.00027	7.6	.023 † .026 ‡	328 † 296 ‡
U / UA distribution wire (buried)	mile	68	"	54	.0014	40.0	.122 † .135 ‡	328 † 296 ‡
Drop wires								
18 gage TP type	Kilofoot	68	Wet	51	.00021	*	.042	......
18 gage TR "	"	68	"	51	.00023	*	.036	......
17 gage BP type	"	68	"	28	.00022	*	.040	......
17 gage BR "	"	68	"	28	.00022	*	.040	......
14 gage HC type	"	68	"	5	.00025	*	.041	......
Miscellaneous wires and cables								
Inside wiring cable—22 gauge	Kilofoot	68	......	37	.00020	*	.025	......
Service cables—22 gage CR type, JR ", LR ", TR "	"	68	......	37 §	.00027 §	*	.020 §	......
AL wire 14 gage	"	68	Wet	5	.00029	*	.033	......
Bridle wire 20 gage	"	68	"	21	.00028	*	.036	......
Duct wire 22 gage; DU station wire 22 gage	"	68	"	33	.00030 ‖	*	.033 ‖	......
GN station wire 22 gage	"	68	"	33	.00030 ‖	*	.048 ‖	......

* Leakage conductance at 1000 cycles is negligible as compared with capacitive susceptance.
† Initial values after one day soaking in water.
‡ Estimated values after 5 to 10 years in ground, depending upon moisture conditions in soil.
§ These values may be applied to both one and two pair cables.
‖ These values are satisfactory for pairs, triples, or quads.

FIG. 27. Primary Distributed Constants of Cables and Miscellaneous Paired Conductor Facilities

Cable Gage	Cable Code	Loading	Propagation Constant at 68 Deg Fahr: Per Mile	Propagation Constant at 68 Deg Fahr: Per Kilofoot	Characteristic Impedance * at 68 Deg Fahr
26	ST AST	NL	.3072 + j .3105	.05818 + j.05881	718 − j706 = 1007 \44.5°
		B-175	.1084 + j .9354	.02053 + j.1772	2204 − j251 = 2218 \6.5°
		B-135	.1207 + j .8223	.02286 + j.1557	1929 − j281 = 1949 \8.3°
		B-88	.1492 + j .6713	.02826 + j.1271	1567 − j344 = 1604 \12.4°
		D-175	.1286 + j .7739	.02436 + j.1466	1848 − j299 = 1872 \9.2°
		D-135	.1434 + j .6824	.02716 + j.1292	1618 − j332 = 1652 \11.6°
		D-88	.1747 + j .5644	.03309 + j.1069	1325 − j403 = 1385 \16.9°
		H-135	.1615 + j .6030	.03059 + j.1142	1440 − j383 = 1490 \14.9°
		H-88	.1940 + j .5049	.03674 + j.09563	1192 − j453 = 1275 \20.8°
		H-44	.2375 + j .4062	.04498 + j.07693	949 − j552 = 1098 \30.2°
		M-135	.1880 + j .5153	.03561 + j.09759	1257 − j460 = 1338 \20.1°
		M-88	.2196 + j .4424	.04159 + j.08379	1057 − j525 = 1180 \26.4°
	BST	NL	.3287 + j .3322	.06225 + j.06292	672 − j660 = 942 \44.5°
		B-175	.1160 + j1.0009	.02197 + j.1896	2060 − j235 = 2073 \6.5°
		B-135	.1292 + j .8799	.02447 + j.1666	1802 − j263 = 1821 \8.3°
		B-88	.1596 + j .7183	.03023 + j.1360	1464 − j322 = 1499 \12.4°
		D-175	.1376 + j .8281	.02606 + j.1568	1727 − j280 = 1750 \9.2°
		D-135	.1534 + j .7302	.02905 + j.1383	1512 − j310 = 1544 \11.6°
		D-88	.1869 + j .6039	.03540 + j.1144	1238 − j376 = 1294 \16.9°
		H-135	.1728 + j .6452	.03273 + j.1222	1346 − j358 = 1393 \14.9°
		H-88	.2076 + j .5403	.03932 + j.1023	1114 − j423 = 1192 \20.8°
		H-44	.2541 + j .4346	.04813 + j.08231	887 − j516 = 1026 \30.2°
		M-135	.2012 + j .5514	.03811 + j.1044	1174 − j430 = 1250 \20.1°
		M-88	.2350 + j .4734	.04451 + j.08966	988 − j490 = 1103 \26.4°
24	M SM ASM CSM	NL	.2467 + j .2513	.04672 + j.04759	558 − j542 = 778 \44.2°
		B-175	.0722 + j .9504	.01367 + j.1800	2155 − j155 = 2161 \4.1°
		B-135	.0794 + j .8344	.01504 + j.1580	1880 − j171 = 1888 \5.2°
		B-88	.0998 + j .6757	.01890 + j.1280	1515 − j216 = 1530 \8.1°
		D-175	.0849 + j .7844	.01608 + j.1486	1800 − j186 = 1810 \5.9°
		D-135	.0941 + j .6887	.01782 + j.1304	1566 − j209 = 1580 \7.6°
		D-88	.1165 + j .5613	.02206 + j.1063	1264 − j257 = 1290 \11.5°
		H-135	.1063 + j .6035	.02013 + j.1143	1386 − j239 = 1407 \9.8°
		H-88	.1309 + j .4945	.02479 + j.09366	1123 − j292 = 1160 \14.6°
		H-44	.1682 + j .3763	.03185 + j.07127	844 − j372 = 922 \23.8°
		M-135	.1254 + j .5066	.02375 + j.09595	1187 − j294 = 1223 \13.9°
		M-88	.1513 + j .4212	.02866 + j.07977	968 − j345 = 1028 \19.6°
	DSM	NL	.2664 + j .2715	.05045 + j.05142	517 − j503 = 721 \44.2°
		B-175	.0780 + j1.0266	.01477 + j.1944	1996 − j143 = 2001 \4.1°
		B-135	.0858 + j .9013	.01625 + j.1707	1741 − j158 = 1748 \5.2°
		B-88	.1078 + j .7298	.02042 + j.1382	1402 − j200 = 1416 \8.1°
		D-175	.0917 + j .8473	.01737 + j.1605	1667 − j172 = 1676 \5.9°
		D-135	.1016 + j .7439	.01924 + j.1409	1450 − j193 = 1463 \7.6°
		D-88	.1258 + j .6063	.02383 + j.1148	1170 − j238 = 1194 \11.5°
		H-135	.1148 + j .6519	.02174 + j.1235	1284 − j222 = 1303 \9.8°
		H-88	.1414 + j .5341	.02678 + j.1012	1039 − j271 = 1074 \14.6°
		H-44	.1817 + j .4065	.03441 + j.07699	781 − j345 = 854 \23.8°
		M-135	.1354 + j .5472	.02564 + j.1036	1099 − j272 = 1132 \13.9°
		M-88	.1634 + j .4550	.03095 + j.08617	897 − j319 = 952 \19.6°
	NM	NL	.2342 + j .2388	.04436 + j.04522	588 − j572 = 820 \44.2°

* Mid-section iterative impedance in cases of loaded facilities.

Fig. 28. Secondary Constants of Exchange Area Cable Facilities at 1000 Cycles per Second—26 and 24 Gage

Cable		Loading	Propagation Constant at 68 Deg Fahr		Characteristic Impedance * at 68 Deg Fahr
Gage	Code		Per Mile	Per Kilofoot	
22	SA	NL	.2065 + j .2134	.03911 + j.04042	416 − j399 = 576 \ 43.8°
	ASA	B-175	.0503 + j1.0155	.00953 + j.1923	2025 − j 92 = 2027 \ 2.6°
	BSA	B-135	.0549 + j .8900	.01040 + j.1686	1762 − j102 = 1765 \ 3.3°
	CSA	B-88	.0689 + j .7177	.01305 + j.1359	1414 − j130 = 1420 \ 5.3°
		D-175	.0583 + j .8365	.01104 + j.1584	1694 − j113 = 1698 \ 3.8°
		D-135	.0647 + j .7325	.01225 + j.1387	1465 − j125 = 1470 \ 4.9°
		D-88	.0808 + j .5922	.01530 + j.1122	1170 − j156 = 1180 \ 7.6°
		H-135	.0729 + j .6402	.01381 + j.1213	1298 − j144 = 1306 \ 6.3°
		H-88	.0907 + j .5185	.01718 + j.09820	1036 − j177 = 1051 \ 9.7°
		H-44	.1199 + j .3796	.02271 + j.07189	748 − j233 = 783 \ 17.3°
		M-135	.0863 + j .5333	.01634 + j.1010	1109 − j178 = 1123 \ 9.1°
		M-88	.1060 + j .4341	.02008 + j.08222	879 − j214 = 905 \ 13.7°
	NA ANA	NL	.1946 + j .2012	.03686 + j.03811	442 − j424 = 612 \ 43.8°
	TA	NL	.1792 + j .1853	.03394 + j.03509	479 − j460 = 664 \ 43.8°
	TS	NL	.1882 + j .1945	.03564 + j.03684	457 − j438 = 633 \ 43.8°
19	BNB	NL	.1446 + j .1551	.02739 + j.02938	295 − j273 = 402 \ 42.8°
	CNB	B-135	.0304 + j .900	.00576 + j.1705	1741 − j 52 = 1742 \ 1.7°
		B-88	.0386 + j .725	.00731 + j.1373	1393 − j 69 = 1395 \ 2.8°
		D-175	.0321 + j .8457	.00608 + j.1602	1676 − j 58 = 1677 \ 2.0°
		D-135	.0349 + j .740	.00661 + j.1402	1448 − j 63 = 1449 \ 2.5°
		D-88	.0439 + j .5957	.00831 + j.1128	1155 − j 81 = 1158 \ 4.0°
		H-135	.0388 + j .6455	.00735 + j.1223	1281 − j 74 = 1283 \ 3.3°
		H-88	.0487 + j .5194	.00922 + j.09837	1013 − j 92 = 1017 \ 5.2°
		H-44	.0645 + j .3701	.01222 + j.07009	713 − j122 = 723 \ 9.7°
		M-88	.0568 + j .4302	.01076 + j.08148	854 − j111 = 861 \ 7.4°
	TB	NL	.1282 + j .1375	.02428 + j.02604	333 − j308 = 453 \ 42.8°
	ANB	B-175	.0254 + j .908	.00481 + j.1720	2237 − j 54 = 2238 \ 1.4°
	DNB	B-135	.0270 + j .795	.00511 + j.1506	1951 − j 61 = 1952 \ 1.8°
		B-88	.0342 + j .641	.00648 + j.1214	1563 − j 76 = 1565 \ 2.8°
		D-175	.0282 + j .7461	.00534 + j.1413	1862 − j 65 = 1863 \ 2.0°
		D-135	.0310 + j .653	.00587 + j.1237	1618 − j 71 = 1620 \ 2.5°
		D-88	.0390 + j .5269	.00739 + j.09979	1292 − j 91 = 1295 \ 4.0°
		H-175	.0315 + j .6507	.00597 + j.1232	1643 − j 75 = 1645 \ 2.6°
		H-135	.0345 + j .5694	.00653 + j.1078	1423 − j 82 = 1425 \ 3.3°
		H-88	.0432 + j .4590	.00818 + j.08693	1132 − j103 = 1137 \ 5.2°
		H-44	.0571 + j .3282	.01081 + j.06216	799 − j138 = 814 \ 9.8°
		M-88	.0505 + j .3796	.00956 + j.07189	948 − j123 = 956 \ 7.4°
16	TH	NL	.0868 + j .1004	.01644 + j.01902	243 − j208 = 320 \ 40.6°
	NN	B-175	.0156 + j .908	.00295 + j.1720	2238 − j 30 = 2238 \ 0.8°
		B-135	.0158 + j .795	.00299 + j.1506	1951 − j 31 = 1951 \ 0.9°
		B-88	.0203 + j .641	.00384 + j.1214	1564 − j 44 = 1565 \ 1.6°
		D-175	.0168 + j .765	.00318 + j.1449	1824 − j 64 = 1825 \ 2.0°
		H-175	.0178 + j .6503	.00337 + j.1232	1648 − j 41 = 1649 \ 1.4°
		H-135	.0188 + j .5687	.00356 + j.1077	1419 − j 42 = 1420 \ 1.7°
		H-88	.0238 + j .4577	.00451 + j.08669	1129 − j 55 = 1130 \ 2.8°
		H-44	.0307 + j .3249	.00581 + j.06153	791 − j 72 = 794 \ 5.2°
		M-88	.0271 + j .3773	.00513 + j.07146	934 − j 75 = 937 \ 4.6°

* Mid-section iterative impedance in case of loaded facilities.

FIG. 29. Secondary Constants of Exchange Area Cable Facilities at 1000 Cycles per Second—22, 19, and 16 Gage

Exchange cables are usually of the non-quadded, paper- or pulp-insulated types, having attenuation losses at 1000 cycles, as shown in Fig. 26. These cables are not generally loaded for subscriber loop use, but are frequently loaded when employed for interoffice trunks, particularly in the larger exchange areas. The primary distributed constants of exchange area cable facilities and of miscellaneous paired conductors for exchange use are shown in Fig. 27. The resistances, inductances and capacitances are in d-c values, but for practical purposes, may be considered equivalent to the 1000-cycle values. The leakage conductances are specifically 1000-cycle values. The secondary constants of of exchange area cable facilities at 1000 cycles, are shown in Figs. 28 and 29.

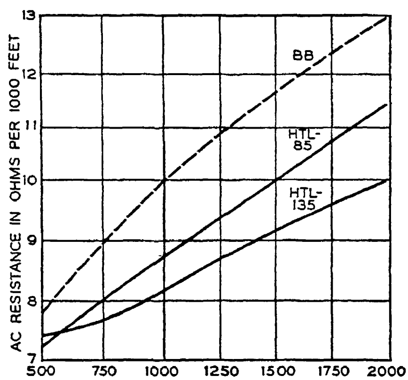


FIG. 30. Comparison of A-c Resistance of BB Grade Iron Wire and Two Types of High Tensile Telephone Line Wire—No. 12 BWG (Copyright 1945 by Indiana Steel and Wire Co.)

OPEN-WIRE FACILITIES for exchange use are principally of iron or high-tensile-strength steel, the steel being used almost exclusively in recent years, owing principally to economy and better service performance. New types of steel have been developed for telephone wire as the result of extensive studies of the materials used, and by means of heat and other treatments applied during manufacture. Some buried wire (such as W.E. Co. U or UA types, loaded or non-loaded) is used in rural areas for short distances, the characteristics of which are given in Fig. 27.

The a-c resistance of BB grade iron wire and of Crapo HTL-85 and HTL-135 high-tensile telephone line wire, over a frequency range of 500 to 2000 cycles, using current of telephonic magnitude, is shown in Fig. 30. The smaller variation in a-c resistance of the Crapo wire (manufactured by Indiana Steel and Wire Co.), as compared to the BB wire, tends to reduce modulation effects on voice currents traveling over the wire, and thus to improve the quality of the transmitted speech. Figure 31 shows comparative breaking strengths between the above three wires, as presented by Indiana Steel and Wire Co. Figure 32 shows the characteristics at 68 deg fahr of 109 (No. 12 BWG) high-strength steel and 134 (No. 10 BWG) steel wire with 0.8-oz zinc coating, 12-in. spacing, and DP insulators.

Copper steel wire (104 mil diameter), the characteristics of which are shown in Table 5, has roughly one-half the attenuation per mile at 1000 cycles of 109 high-strength steel wire (Fig. 32). Thus, this wire may be used advantageously in place of high-strength steel wire where the lower loss is required to meet exchange transmission standards. Copper steel wire (0.081 mil diameter and 40 per cent conductivity) having per pair mile an attenuation loss of 0.22 db (wet) and resistance of 42.8 ohms at 68 deg fahr may also be used where applicable.

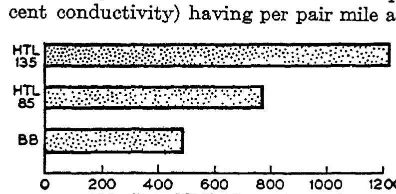


FIG. 31. Comparative Breaking Strength of BB Grade Iron Wire and Two Types of High Tensile Telephone Line Wire—No. 12 BWG (Copyright by Indiana Steel and Wire Co.)

Buried wire (paired, insulated, such as W.E. Co. U or UA type) is sometimes used, loaded or non-loaded, in place of or as an extension of open-wire rural plant, depending on economies.

CARRIER CHANNELS superimposed on rural power lines are now under trial operation to determine the practicability and requirements for thus serving rural subscribers in locations not at present served by telephone lines. The equipment, developed by the Bell System and designated as the M-1 carrier telephone system, is being made in quantities by the Western Electric Co., the privilege of producing it being extended to other manufacturers.

The first two systems, installed in 1945 for trial at Jonesboro, Ark., and Selma, Ala., provide rural service to four subscribers over an 11-mile section of 7200-volt, multigrounded neutral, single-phase power line out of Jonesboro, and to four subscribers over a 14-mile section of 6900-volt, multigrounded neutral, single-phase power line out of Selma.

The principal elements of the rural power line carrier system, as now developed, are shown in Fig. 33. Single channel operation is employed, using for the trials frequencies

0.8-oz. Zinc Coating, 12-in. Spacing, 68 Deg Fahr
DP Insulators

Frequency, kc	Attenuation, db/mi		Phase Shift, radians/mi		Characteristic Impedance	
	Dry	Wet	Dry	Wet	Dry	Wet
109 Side Circuits						
0.3	.181	.207	.028	.026	$1754 - j1314 = 2192\overline{\backslash 36.83^\circ}$	$1818 - j1122 = 2136\overline{\backslash 31.68^\circ}$
1.0	.289	.313	.067	.068	$1279 - j\,629 = 1425\overline{\backslash 26.18^\circ}$	$1274 - j\,578 = 1399\overline{\backslash 24.40^\circ}$
2.0	.407	.436	.121	.095	$1142 - j\,443 = 1225\overline{\backslash 21.20^\circ}$	$1130 - j\,413 = 1203\overline{\backslash 20.08^\circ}$
3.0	.528	.560	.167	.169	$1054 - j\,383 = 1121\overline{\backslash 19.97^\circ}$	$1042 - j\,360 = 1102\overline{\backslash 19.07^\circ}$
4.0	.623	.660	.212	.215	$1004 - j\,339 = 1060\overline{\backslash 18.65^\circ}$	$991 - j\,320 = 1041\overline{\backslash 17.90^\circ}$
109 Phantoms (of Non-pole Pairs)						
0.3	.166	.195	.026	.025	$969 - j\,713 = 1203\overline{\backslash 36.35^\circ}$	$1006 - j\,588 = 1165\overline{\backslash 30.30^\circ}$
1.0	.259	.286	.064	.065	$721 - j\,334 = 795\overline{\backslash 24.85^\circ}$	$717 - j\,300 = 777\overline{\backslash 22.70^\circ}$
2.0	.366	.396	.115	.116	$642 - j\,237 = 684\overline{\backslash 20.27^\circ}$	$635 - j\,215 = 670\overline{\backslash 18.70^\circ}$
3.0	.466	.502	.160	.162	$598 - j\,200 = 631\overline{\backslash 18.50^\circ}$	$589 - j\,185 = 617\overline{\backslash 17.45^\circ}$
4.0	.548	.588	.204	.207	$572 - j\,176 = 598\overline{\backslash 17.10^\circ}$	$563 - j\,164 = 586\overline{\backslash 16.25^\circ}$
134 Side Circuits						
0.3	.136	.159	.026	.025	$1563 - j\,945 = 1826\overline{\backslash 31.17^\circ}$	$1599 - j\,788 = 1783\overline{\backslash 26.23^\circ}$
1.0	.250	.273	.069	.070	$1252 - j\,520 = 1356\overline{\backslash 22.55^\circ}$	$1245 - j\,473 = 1332\overline{\backslash 20.80^\circ}$
2.0	.389	.416	.120	.121	$1084 - j\,404 = 1157\overline{\backslash 20.43^\circ}$	$1073 - j\,377 = 1137\overline{\backslash 19.36^\circ}$
3.0	.502	.533	.167	.168	$1004 - j\,348 = 1063\overline{\backslash 19.12^\circ}$	$992 - j\,328 = 1045\overline{\backslash 18.30^\circ}$
4.0	.599	.632	.209	.212	$945 - j\,311 = 995\overline{\backslash 18.22^\circ}$	$933 - j\,294 = 978\overline{\backslash 17.48^\circ}$
134 Phantom Circuits (of Non-pole Pairs)						
0.3	.123	.149	.024	.023	$883 - j\,524 = 1027\overline{\backslash 30.68^\circ}$	$902 - j\,415 = 993\overline{\backslash 24.70^\circ}$
1.0	.220	.246	.064	.064	$711 - j\,281 = 765\overline{\backslash 21.57^\circ}$	$706 - j\,249 = 749\overline{\backslash 19.43^\circ}$
2.0	.341	.371	.112	.114	$625 - j\,219 = 662\overline{\backslash 19.32^\circ}$	$617 - j\,199 = 648\overline{\backslash 17.88^\circ}$
3.0	.436	.470	.156	.158	$579 - j\,185 = 608\overline{\backslash 17.72^\circ}$	$570 - j\,172 = 595\overline{\backslash 16.79^\circ}$
4.0	.516	.554	.197	.200	$548 - j\,165 = 572\overline{\backslash 16.75^\circ}$	$539 - j\,153 = 560\overline{\backslash 15.85^\circ}$

FIG. 32. Characteristics of 109 (No. 12 BWG) High-strength Steel and 134 (No. 10 BWG) Steel Wire

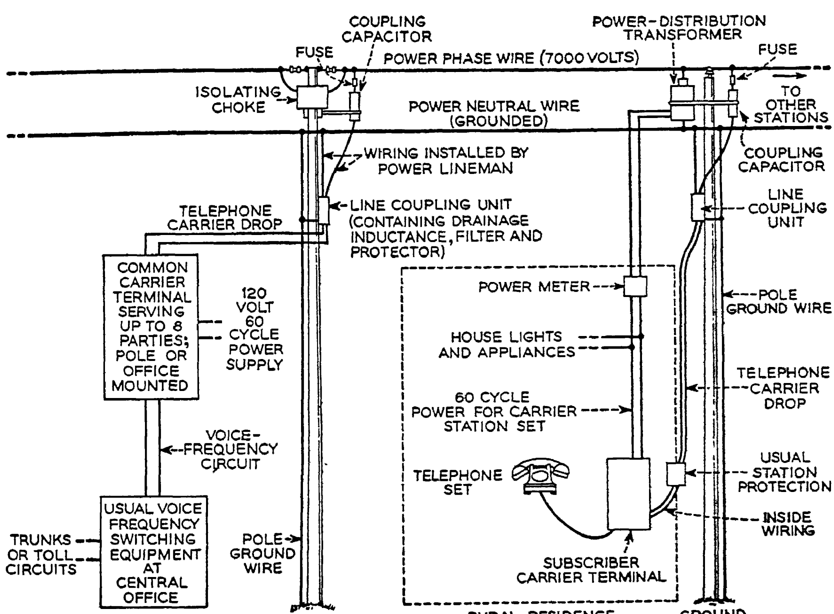


FIG. 33. Principal Elements of Rural Power Line Carrier Telephone System (Courtesy Bell System)

of 165 kc central office calling subscriber and subscriber receiving, 195 kc subscriber calling central office and subscriber talking out through central office, and 185 kc one subscriber calling another subscriber on the same line but through the central-office carrier terminal equipment. The M-1 system is designed, however, for double sideband carrier transmitted amplitude modulation, with as many as six channels, each serving eight subscribers and each using three frequencies. These frequencies are different within each channel and for each channel, being selected for transmission from the common terminal to the stations within the range 155 to 230 kc and for transmission from the stations to the common terminal within the range 290 to 450 kc.

Figures 34 and 35 show block diagrams of the common carrier terminal at the central-office end of the system and of the subscriber carrier terminal (station set), respectively, of the types used in the trial tests.

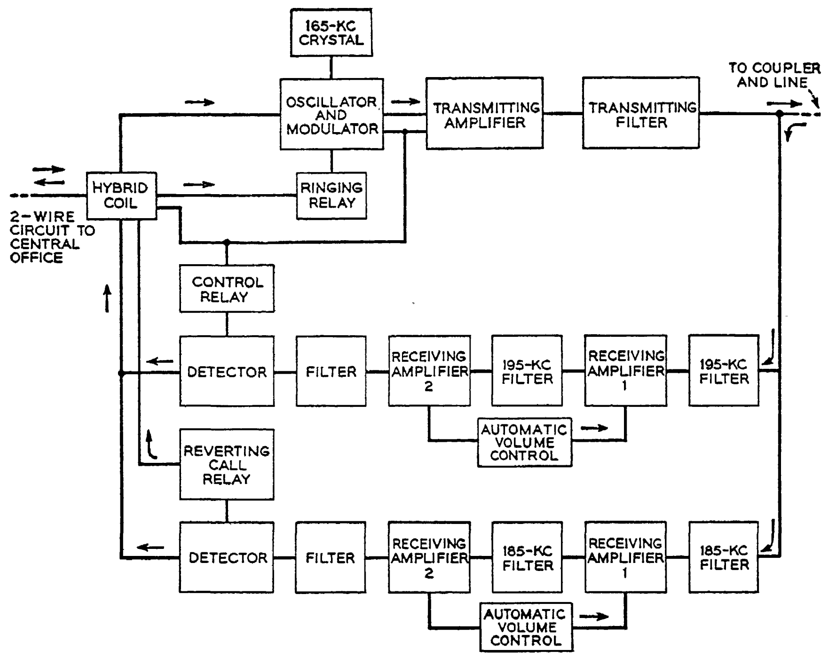


FIG. 34. Block Diagram of Common Carrier Terminal at the Central-office End of a Rural Power Line Carrier Telephone System (Courtesy Bell System)

The carrier terminals are connected to the power line through a 0.002-μf capacitor (8700-volt rating for the trials), and a line coupling unit. This unit has a drainage inductance coil, filter, and protector. All branches of the power network not used for carrier transmission have inserted in the primary phase conductor an isolating choke for the main power circuit and a tap choke for branches from the main circuit. Transmission chokes are used in branches from the main power circuit, where the branch is being used for carrier, in order to reduce the bridging loss of the branch to through transmission over the main line. The common carrier terminal is designed to terminate the power line in about 500 ohms impedance for the carrier frequencies used, as are the coupling capacitors and line coupling units at subscriber stations or at the end of a power line tap where no subscriber station exists.

Each carrier terminal requires continuous 110–120 volt, 60-cycle a-c power supply, and the unmodulated carrier power delivered to the line coupling unit from each terminal is about 1 watt. The subscriber terminal requires about 8 watts of standby and 30 watts of operating power.

The common carrier terminal connects to the manual switchboard or dial unit over a regular two-wire voice frequency circuit. Divided code ringing is provided using interrupted carrier current in proper time sequence, or bridged code ringing can be used if required.

Where the power circuit is not of the single-phase, multigrounded neutral type or it is desired to operate more than one system on the same power line but in different sections, special consideration of the factors involved will be required.

Preliminary data indicate that the average transmission loss between the common terminal and the most distant station at the operating carrier frequencies of the M-1 system, including subscriber coupling unit bridging and transmission choke losses, will be about 2.0 to 2.5 db per mile. The overall loss between the common terminal and any station should not exceed 35 to 40 db (15 to 18 miles of line). Under low atmospheric static conditions, the carrier system can be adjusted to provide about the same effective transmission $(T + R)/2$ from any station to the central office as would be provided from a regular voice-frequency station having the latest-type antisidetone, local battery talking set (W.E. Co. FIA-AST-LBT-2 cells).

The M-1 carrier system may also be adapted to telephone wire lines, where additional telephone circuits may thus be provided more economically than by other methods. This usage requires further study.

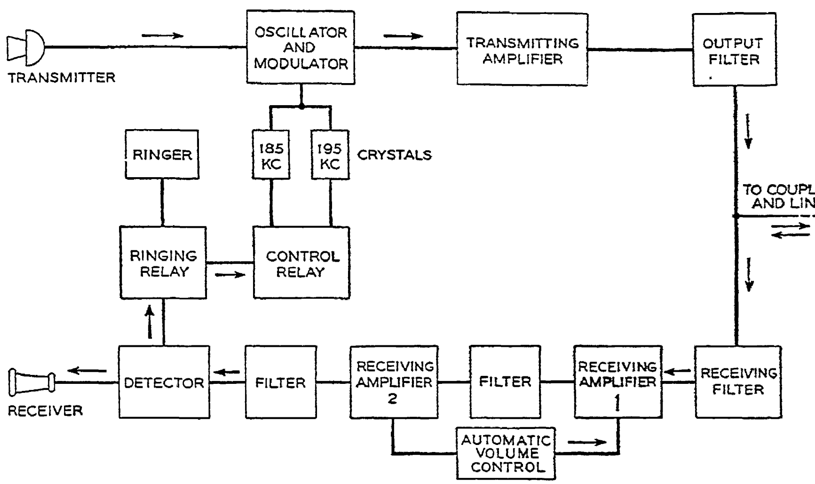


Fig. 35. Block Diagram of Subscriber Station Set of a Rural Power Line Carrier Telephone System (Courtesy Bell System)

In operation, the system responds as though the connecting circuit between the central office and subscriber were the usual wire line. On revertive calls (one subscriber calling another on the same circuit or channel), the calling subscriber places the call to the called subscriber in the usual manner (through the operator or equipment) and then puts his handset on the cradle. When the ringing signal, which the calling subscriber hears, is stopped, indicating that the called subscriber has answered, the calling subscriber removes his handset from the cradle and talks. The conversation is carried on through the common terminal, and operator supervision is provided if the office is of the manual type.

Radio channels for rural telephone service are under trial test, as discussed in article 8 of this section. It is expected that suitable equipment will be developed, based on these tests, which will permit utilizing radio channels for this service on a commercial basis.

SUBSCRIBER LOOP DESIGN. The design procedure and considerations in engineering subscriber loop plant are briefly summarized (for a single cable route) as follows:

(*a*) Determine the most economical gage requirements for the ultimate area to be served by the various complements of cable, considering both transmission and signaling design limits, using the effective subscriber loop loss curves, if available.

(*b*) Examine the possibilities of obtaining further cable economies by reducing bridged tap or distribution cable losses which were assumed in determining the design limits.

(*c*) Examine the possibilities of utilizing any existing plant with the new plant being designed, without exceeding the design limits.

(*d*) If the loop loss or signaling limits are exceeded in existing plant, determine the most economical plan of meeting the limits, such as using special sets, reducing bridged cable losses, loading the longer cable pairs, or choosing larger-gage cable or long line circuits.

(e) If some margin of loop loss and loop resistance results from the use of existing plant in connection with the plant being designed, study the advisability of smaller-gage pairs than indicated under (a) above, taking into consideration the permanence of the existing facilities and the relation of the proposed plant to the ultimate plan.

(f) Combinations of gages are frequently practicable, particularly since the trend is toward the finer-gage cable adjacent to the central office.

(g) Composite (more than one gage) cables are installed extensively, the larger gages serving the outlying areas.

(h) Loading on long loops is sometimes employed to meet loop loss limits. H-44 or H-88 loading on 19-, 22-, or 24-gage cable results in substantial loss reductions over non-loaded loops, which might permit smaller-gage conductors for long loops. Bridged taps on loaded pairs must be limited to avoid serious impairments.

(i) Signaling limits may economically be extended, in some cases by modifications or readjustments of central-office equipment, larger-gage conductors or long line units for the longer loops, or occasionally the use of two pairs in parallel may be justified to reduce loop resistance.

(j) Loop and trunk plant design is based on the most efficient types of subscriber sets available or, in some cases, anticipated within a relatively short period of time.

(k) Special exchange lines, such as private or PBX tie and foreign exchange lines, conference and bridging arrangements, and one-way speech networks, require special design work to meet their particular needs.

The *zoning* of subscriber sets is a procedure that provides for the introduction of station apparatus in such a way as to obtain the desired grade of transmission at the lowest practicable overall cost. Owing to the different efficiencies of the various types of station apparatus in service and under continuous development, it is necessary to insure that this apparatus will be installed as required, so that (1) the cost of the outside plant will be the minimum, (2) transmission will be satisfactory for the outside plant design, and (3) station apparatus costs will be minimized by obtaining a reasonable service life for the older apparatus, avoiding as far as possible premature replacements of existing apparatus and providing an orderly program for the introduction of new station apparatus. Thus, exchange areas, as required, are divided into zones within which only certain types of station apparatus may be used.

TRUNK PLANT DESIGN is based on the transmission loss limit assigned to each group of trunks, as determined from the loop and trunk study. The *design limit* is obtained by deducting from the overall permissible trunk loss (1) terminal junction losses, due to characteristics of the proposed trunks, which differ from the reference trunk; (2) intermediate junction losses, due to a trunk being composed of different types of facilities; (3) losses due to loading irregularities; and (4) equipment and office losses at intermediate points between the trunk terminals. Data have been prepared for the Bell System showing both terminal and intermediate junction losses under the usual conditions encountered in practice. In some cases, under certain conditions, these losses are actually negative (transmission gains).

The design limits for the trunks, having been determined for the various conditions under which the given group of trunks will operate, represent the *permissible effective trunk losses*. The type of trunk is then selected which meets the design limits and has least outside plant costs, taking into consideration existing trunk plant, future trunk facility needs over the route involved, and any other factors that may have a bearing on the selection of the type of trunk. Figure 36 shows *effective losses* for non-quadded exchange area cable trunks.

The signaling limits for trunks are based on signaling, pulsing, and supervision requirements, which vary materially for different types of offices and associated terminal equipment. These limits, as well as leakage requirements, are usually indicated on the standard drawings for each type of central-office circuit, which may be different for the two ends of the given trunk. The lower value, if there is a difference, will be controlling in determining the signaling limit.

Overall trunk resistance for signaling purposes is usually computed from unit values for temperatures of 68 deg fahr. For underground trunk plant the resistance change with temperature will be about 3 per cent maximum, which may usually be disregarded. For aerial cable, the change may be as much as 10 to 12 per cent, and this variation of resistance with temperature change should be considered in the signaling design limits.

With the larger-gage trunk facilities generally used in the past, the transmission design limit usually controlled the selection of the type of trunk facilities, but with the higher permissible interlocal trunk losses in recent years (resulting from allocation to trunks of part of the new instrument gains) and the trend toward higher supervision limits with current types of dial equipment, the field of use of the finer-gage cables has been extended, so that signaling limits may be controlling in some cases.

Cable		26 AST	24 ASM	22 BSA	19 CNB	19 DNB	16 TH	13 TJ
Loading	Spacing ft	Decibels per Mile at 68 Deg Fahr						
NL	....	3.20	2.57	2.15	1.51	1.34	0.91	0.60
M-88	9000	....	1.35	0.95	0.51	0.46	.26	.16
M-135	9000	....	1.22	0.79	0.42	0.37	.21	....
M-175	9000	....	....	....	....	0.34	.17	.12
H-44	6000	....	1.48	1.05	0.57	0.51	.27	....
H-88	6000	....	1.15	0.80	0.43	0.39	.21	....
H-135	6000	....	....	0.64	0.35	0.31	.17	....
H-175	6000	....	....	....	0.32	0.28	.16	.11
D-175	4500	....	....	0.52	0.29	0.26	.15	....
B-88	3000	....	0.87	0.61	0.35	0.31	.18	....
B-135	3000	....	....	0.49	0.27	0.25	.14	....
B-175	3000	....	....	....	0.26	0.23	.14	....
		Decibels per Mile at 110 Deg Fahr						
NL	....	3.33	2.68	2.24	1.57	1.39	0.94	0.63
M-88	9000	....	1.47	1.03	0.56	0.50	.28	.17
M-135	9000	....	1.33	0.86	0.46	0.40	.23	....
M-175	9000	....	....	....	....	0.37	.18	.13
H-44	6000	....	1.61	1.15	0.62	0.55	.30	....
H-88	6000	....	1.25	0.87	0.47	0.42	.22	....
H-135	6000	....	....	0.70	0.38	0.34	.19	....
H-175	6000	....	....	....	0.35	0.31	.18	.12
D-175	4500	....	....	0.56	0.31	0.28	.17	....
B-88	3000	....	0.94	0.66	0.38	0.33	.20	....
B-135	3000	....	....	0.53	0.30	0.27	.15	....
B-175	3000	....	....	....	0.28	0.25	.15	....

Note: The resultant trunk loss should not be used closer than to nearest 0.1 db.

FIG. 36. Effective Trunk Losses—Exchange Area Cable Trunks

PROGRAM SERVICE

18. PROGRAM SERVICE

In general, program service, as furnished by the telephone companies, consists mainly of providing suitable wire or carrier facilities to the broadcasting companies for the transmission of program material, which usually originates in broadcasting studios or other locations and which is broadcast from radio transmitters to the public.

Programs, being of a varied nature, from the finest orchestral music to ordinary speech, require different grades of telephone facilities to meet the broadcaster's requirements. For this reason the telephone companies have developed and have made available for broadcasting purposes several rather broad classifications of facilities, as shown in Table 1.

SERVICE REQUIREMENTS, being more exacting for high-quality than the lower-quality program circuits, involve the control of (1) transmission levels and losses, (2) frequency, delay, and phase distortion, and (3) noise and crosstalk. See Section 12 for critical cutoff points in the frequency ranges of various musical instruments and speech.

Transmission levels and losses in open wire and cable are maintained as specified, by special program amplifiers, the latest types having a flat gain characteristic with a maximum gain of about 30 to 40 db and outputs of about 10 to 20 db above reference volume,

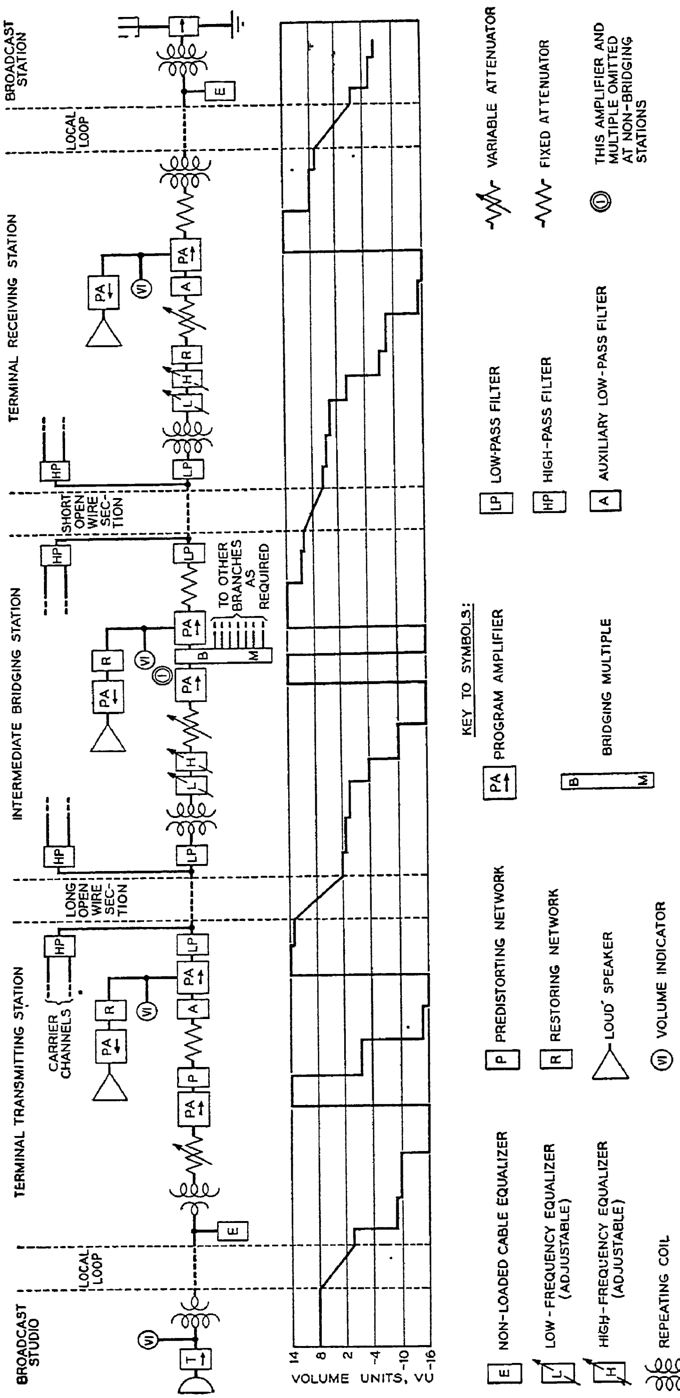


Fig. 1. Typical Layout and Level Diagram for a Wide-band Open-wire Program Circuit (Courtesy Bell System)

depending on the type of amplifier. The levels must be high enough to provide a satisfactory signal-to-noise ratio and must also be coordinated with the levels of adjacent circuits or systems, to avoid causing interference in them. A variable attenuator, having a range of 0 to 32 db, controls the amplifier gain. Figure 1 shows a typical layout and level diagram for a wide-band (8000 cycles) open-wire program system.

Attenuation variations in the line over the frequency band (frequency distortion) are compensated for by low- and high-frequency equalizers of the constant-resistance type, which correct for variations in the preceding line section, to meet requirements.

A repeating coil (impedance ratio, line to drop, of 1 : 1.15) is provided at each repeater point between the incoming line and line equalizing apparatus, which latter is on the line side of the line amplifier. This coil gives the proper termination to the line and insulates the line from the terminal equipment against noise and for protection. Suitable repeating coils are also provided at both ends of the local loops for similar reasons. A non-loaded cable equalizer is provided for the local cable loops.

Special filters are required for facilities which transmit both program and carrier frequencies, in order to separate the two frequency bands at the line terminations and direct them into their proper equipments. Line filters, designed for 5000-cycle program systems, introduce slight delay distortion in the program frequency bands. If the number of line sections in tandem equipped with these filters is about eight or more, a *delay equalizer* for each two filters is necessary to maintain the time of propagation of all frequencies in the program band within satisfactory limits (not to exceed about 0.3 millisecond difference between maximum and minimum delays for a range of 500 to 5000 cycles). An auxiliary low-pass filter may be provided, if required, for 5000-cycle systems, to effect further discrimination against high-frequency interference from the line.

Table 1

Classification of Facility	Approximate Frequency Band in Cycles
Intercity Circuits	
1. High quality	100– 5,000 50– 8,000
2. Medium quality	200– 3,500
3. Speech only	300– 2,500
Metropolitan Area Circuits	
1. Studio-transmitter circuits for AM stations	50– 8,000 *
2. Studio-transmitter circuits for FM stations	50–15,000
3. Network loops—between studio and point of connection (toll office) with intercity network channels	†
4. Pick-up circuits—between points of program origin and studio or point of connection with intercity channels	†

* Band may be extended to higher frequencies if specifically requested by the customer.

† Equalized at request of the customer for band width desired (usually 50–8,000 cycles).

An auxiliary low-pass filter may be used on 8000-cycle program circuits to supplement the low-pass filter of the carrier line filter set where further discrimination is needed.

Noise must be limited in all program circuits, so that it will not interfere appreciably with the quality of the broadcast.

Predistorting and restoring networks are employed in open-wire program systems, particularly the 8000-cycle system, to minimize high-frequency noise from nearby carrier systems. Predistortion is accomplished by introducing, at the sending end of the circuit, a network which effectively raises in volume the currents above 1000 cycles to a higher level than normal for line transmission, thus increasing the signal-to-noise ratio at these frequencies. Since the power at the higher frequencies is relatively small, the amplifiers are not overloaded by this procedure. The restoring network at the receiving end of the circuit restores the predistorted currents to their original amplitude and phase relation. The net reduction in high-frequency interference is equal to the relative losses introduced by the restoring network at the frequencies restored. Figure 2 shows the characteristics of these networks, which are of the lattice type and are composed of combinations of inductance, capacitance, and resistance elements.

Crosstalk also must be controlled. Methods of limiting crosstalk include such items as proper selection of facilities, maintenance of proper levels, and avoidance of the adverse effects of circuit irregularities which may develop from time to time.

PROGRAM FACILITIES may be of open wire or cable, assigned for program use in the program frequency band, or such facilities may consist of single-sideband transmission over cable carrier systems, using three channels of a twelve-channel unit, or over other carrier system channels.

Open-wire facilities for *long-haul* broadcasting usually consist of 165-mil hard-drawn copper-wire circuits, although 128- or 104-mil wire is frequently used, where available and appropriate, but generally for the shorter-haul circuits. Preference is given to 8-in. non-phantomed pairs.

Amplifiers are spaced about twice as frequently (50 to 150 miles) for 104 circuits as for 165 circuits. Open wire, being more subject to noise and crosstalk than cable or carrier, must be carefully selected and maintained in order to provide facilities of suitable quality.

Cable pairs (non-quadded) may be employed for *long-haul* program service. The latest type of cable system for this purpose is the 16-gage B-22 loaded system, which is satisfactory for a band of 35 to 8000 cycles.

The 16-gage B-22 system operates over one-way transmission paths. The non-quadded cable pairs are loaded at a 3000-ft nominal spacing with 22-millihenry coils. The attenuation-frequency characteristics of this facility are shown in Fig. 19, article 16, and the impedance, cutoff, and velocity values are shown in Fig. 18, article 16. Amplifiers and

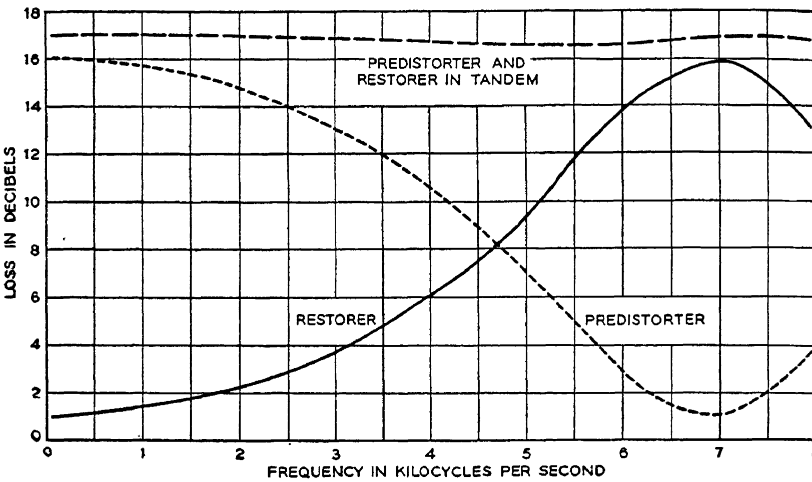


FIG. 2. Loss-frequency Characteristics of Predistorting and Restoring Networks—Open-wire Program System (Courtesy Bell System)

associated apparatus are located at about 50-mile intervals. Attenuation and delay equalizers, as required, are associated with each line amplifier input circuit. A supplementary adjustable equalizer is usually required in each repeater regulator section, because of the variations in capacitance and conductance of the different cable sections through which the program circuit passes. The application of this system for new installations may be limited somewhat in the future by the newer developments in *single-sideband carrier program transmission.*

High-quality program circuits may be operated over cable carrier systems, using single-sideband transmission and three channels of the system. The single-sideband system is designed to operate over as many as ten carrier links in tandem. Figure 3 shows a schematic of a single-sideband program terminal arranged for transmitting over cable carrier.

Medium-quality program circuits may be assigned to 19- or 16-gage H-44 or B-88 cable side circuits, but non-linear and delay distortion in the repeaters and loaded pairs and transmission variations, particularly at high and low frequencies, limit the length of these facilities to about 300 miles for H-44 and B-88 loaded facilities. Channels of the cable carrier system may be used for medium-quality service if not more than one link of these carrier systems is used. Non-loaded cable is limited to relatively short lengths.

METROPOLITAN-AREA PROGRAM FACILITIES, as indicated in Table 1, are designed to provide (1) studio to transmitter circuits for both a-m (amplitude-modulation) and f-m (frequency-modulation) broadcasting stations, (2) network loops between the studio and the point of connection (usually the telephone company toll office) with intercity network channels, and (3) pick-up circuits between the point of program origin and the studio or the point of connection with intercity network channels.

The types of line facilities used may consist of non-loaded exchange or toll cable pairs of various available gages, loaded exchange or toll cable pairs (where loading is required for transmission reasons), or open-wire pairs. Metropolitan-area program circuits, to a large extent, employ non-loaded exchange cable pairs, and in some cases open-wire pairs, with or without intermediate amplifiers.

Equalization may or may not be required, depending on the band width to be transmitted, length and gage of facilities involved, and the transmission deviation permissible over the frequency band.

For studio to transmitter circuits, serving a-m stations, a relatively flat transmission-frequency characteristic may be obtained between 50 and 8000 cycles, using standard equalizer equipment, for sections of non-loaded cable not exceeding about 21.5 miles of 16-, 10.0 miles of 19-, 6.5 miles of 22-, 5.0 miles of 24-, or 4.2 miles of 26-gage cable conductors. Longer lengths of cable may be divided into sections not exceeding the above lengths, which are then treated individually. It is desirable, from a transmission stand-point, to employ a uniform type of facility for program circuits when available.

For studio to transmitter circuits, serving f-m stations, standard equalizer arrangements provide a satisfactory transmission characteristic over metropolitan-area circuit lengths usually encountered. In one instance, such a circuit, consisting of 24 miles of 19- and 22-gage cable pair (non-loaded), with three intermediate amplifiers, was equalized to limit the overall transmission variation to 1.9 db for a frequency band of 30 to 15,000 cycles. Wider bands, extending to 18 or 20 kc, have been provided in special cases when requested by the customer, using special equalizing methods.

Carrier loaded (C 4.1 or C 4.8 loading) cable facilities when available may be equalized for metropolitan-area circuits, serving either a-m or f-m stations. The B 22 facilities are suitable for frequencies up to 8000 cycles for any length of circuit likely to be encountered, using amplifiers and equalization as required.

SPECIAL FEATURES developed for program systems include:

(*a*) Bridging arrangements, in which provision is made for connecting one or more branch program circuits to the main program circuit at a given point.

(*b*) Monitoring, in which attendants may listen in and supervise programs (to insure satisfactory operation) at designated points on the broadcast network.

(*c*) Reversals, in which provision is made for two-way transmission over the same network facilities, by reversing the direction of transmission of all connected one-way apparatus at will, either manually or automatically, and under control of the telephone company or the customer.

(*d*) Order wire and talking arrangements, which permit the various control points and attendants to converse readily regarding network operations without interfering with the broadcast facilities.

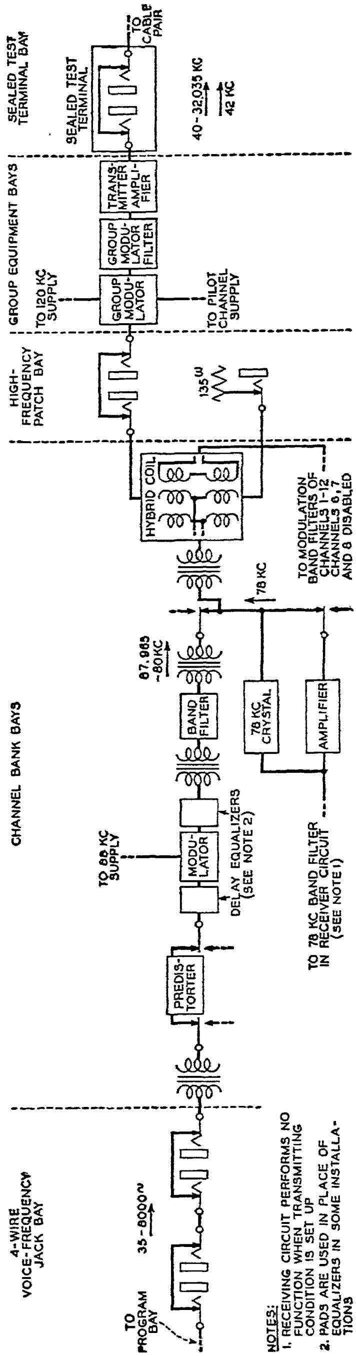


FIG. 3. Single-sideband Program Terminal Arranged for Transmitting over Cable Carrier System (Courtesy Bell System)

SUBSCRIBER STATIONS

19. SUBSTATION EQUIPMENT

SUBSCRIBER STATION EQUIPMENT consists of a large number of types and designs of subscriber telephone sets and auxiliary apparatus, essential to the furnishing of a complete telephone service in the most economical and satisfactory manner.

The basic subscriber telephone set is assembled in a number of designs to meet the needs of different services and the subscriber's convenience, but the operating principle is primarily the same for all these sets. The basic set consists principally of a transmitter, receiver, induction coil, condensers, ringer, and spring assembly (switchhook), suitably mounted in a metal or plastic housing. The set will also have a dial if it is connected to a dial exchange or unit, and for magneto telephones a dry-cell battery for transmitter current and a magneto generator for signaling must be provided.

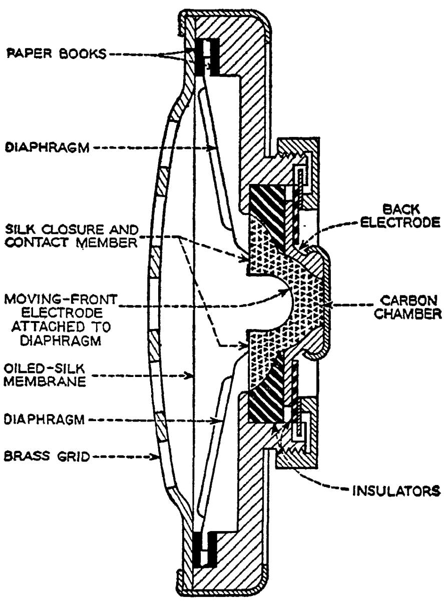


FIG. 1. Cross-sectional View of Modern Non-positional Transmitter Unit (Courtesy Bell System)

THE TELEPHONE TRANSMITTER is designed to receive airborne sound waves of various frequencies and convert them into electrical waves of similar frequencies for transmittal over a telephone circuit. It is essentially a device which makes it possible for relatively weak sound energy to control electrical energy of greater average strength.

The modern transmitter unit employs an insulated, spherical-shaped carbon chamber holding the carbon granules, and a very light metal conical diaphragm, with a dome-shaped center, which is positioned in the carbon chamber in such a way as to hold the granules in the chamber. The dome and chamber are the front and back electrodes, respectively, of the unit. Figure 1 shows a cross-sectional view of the Western Electric Co. No. F1 unit. The electrode surface area in contact with the carbon granules is the same, regardless of the position in which the transmitter is used. The No. F1 unit has a resistance (new) of about 30 to 40 ohms and is applicable to any type of subscriber telephone set of Western Electric Co. make. Transmitter current should not exceed about 100 milliamperes for normal usage and life. A small condenser is usually connected directly across

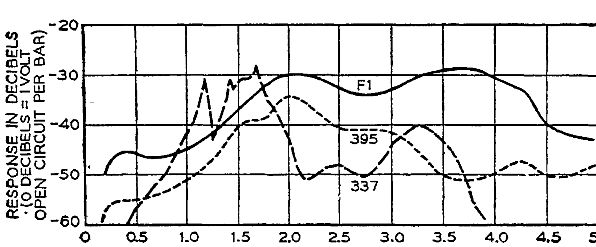


FIG. 2. Transmitter Response Characteristics (Courtesy Bell System)

the transmitter electrodes of the F1 unit, when used in handsets, to minimize packing (welding together) of carbon granules as the result of arcing in the carbon chamber when the transmitter current is rapidly interrupted by switchhook flashing or other operations.

The response of all makes of the modern non-positional transmitter is greatly improved over the older types, as shown for Western Electric Co. instruments in Fig. 2.

THE TELEPHONE RECEIVER is designed to receive electrical waves of various frequencies and convert them into sound waves of similar frequencies for the listener's ear. It is essentially a reconverter from electric to sound energy (within its frequency range) and is a necessary complement of the transmitter in the transmission of speech.

The modern receiver is generally of the unit or capsule design, which is applicable to various types of telephone receivers and assembled sets as made by the respective manufacturers. One make (W.E. Co. HA1) consists principally of a bipolar permanent magnet, with the parts assembled in a zinc alloy frame. The diaphragm of Permendur is seated on a ring projection of the frame, just above the magnet, with an air chamber of definite volume behind it. This chamber has a small outlet hole covered with a silk disk of specified acoustic impedance. The diaphragm also has an air chamber of definite volume between it and the receiver cap, which latter has six holes of definite length and area. The unclamped diaphragm thus rests between two air chambers of specified volumes and outlet impedances, and variations in receiver efficiency with temperature changes as well as diaphragm freezing to the pole pieces are practically eliminated.

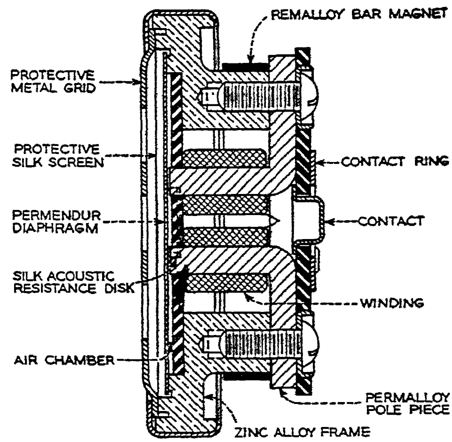


Fig. 3. Cross-sectional View of Modern Receiver Unit (Courtesy Bell System)

The HA1 receiver unit has a working impedance at 1000 cycles of about 140 ohms with a positive angle of 60°. Figure 3 shows a cross-sectional view of this type of receiver unit, and Fig. 4 shows its response characteristic as well as that of the older type W.E. Co. No. 557 receiver.

The HA1 receiver is affected adversely by d-c flow of either polarity, which for 100 milliamperes amounts to about 4.5 and 6.0 db loss in volume efficiency for the opposing and aiding directions, respectively, of current flow.

The anti-sidetone type induction coil is now generally employed, in place of the older sidetone type. Figure 5 shows schematic diagrams of both sidetone and anti-sidetone station circuits and the direction of instantaneous current flow for both the transmitting and receiving conditions.

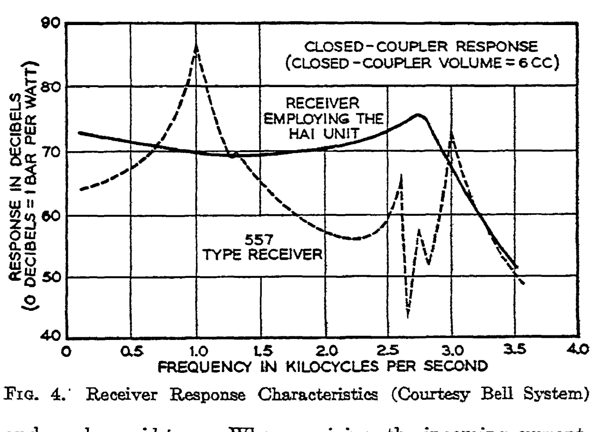


Fig. 4. Receiver Response Characteristics (Courtesy Bell System)

In the *sidetone connection* (Fig. 5, circuit 1), the speech currents produced in the transmitter divide between the A and B windings (solid arrows). The current in each of these windings induces a voltage in the other winding (dashed arrows). The two currents in A combine to flow out over the line, and the two currents in B combine to flow through the receiver and produce *sidetone*. When receiving, the incoming current divides between the transmitter (T) and the receiver (R) branches (solid arrows). The current in A induces a voltage in B, which establishes an opposing current in B (dashed arrow), but the incoming line current in B is larger, so that the resultant current through R is sufficient to actuate the receiver diaphragm.

The *sidetone reduction connection* of the sidetone coil employed on short loops having the sidetone type of subscriber set is shown in Fig. 5, circuit 2. With this connection the transmitter current cannot flow directly through the receiver (R) branch, but some induced current does flow through R because of the inductive coupling between windings A and B. Since transmitter current does not flow directly through B, this winding does not induce a voltage in A, such as occurred in the sidetone connection, and the transmitter line current is not thus aided, resulting in lower transmitting efficiency. In receiving, the incoming line current does not enter the B winding directly but flows through winding A, establishing an induced voltage from A to B, which causes received current to flow through R. Thus, the induced R current is not opposed by line current flowing directly, as was the case in the sidetone connection, resulting in slightly higher receiving efficiency.

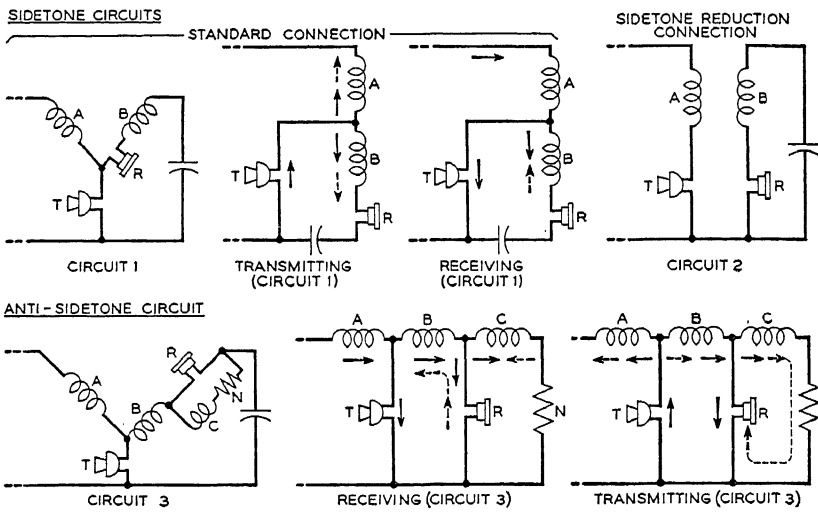


FIG. 5. Sidetone and Anti-sidetone Station Circuits (Courtesy Bell System)

In the *anti-sidetone connection* (Fig. 5, circuit 3), a third winding of relatively high resistance, so that the winding has both an inductance (C) and resistance (N) component, has been added to the sidetone circuit.

In receiving, the incoming current divides between the transmitter (T) branch and winding B (solid arrows). The current in B divides between the receiver (R) branch and the winding C (solid arrows). The currents in A and B induce voltages in C, and, by properly proportioning the coil winding relations, the current in C (dashed arrow) resulting from these combined induced voltages is opposite and about equal to the current flowing directly in C. The C winding thus has no appreciable effect on *receiving volume*, and the *receiving efficiency* of both the sidetone and anti-sidetone circuits is about the same. However, owing to sidetone suppression, the *effective receiving losses* for the anti-sidetone circuit are substantially less.

In transmitting, the transmitter (T) output current divides between the windings A and B, and the current entering B divides between the receiver (R) branch and the C winding (solid arrows). These currents in A and B induce voltages in B and A, respectively, resulting in an induced current flowing in an aiding direction in the line and in B (dashed arrows). The induced current in B divides between R and winding C (short dashed arrow). Also, there is induced in C a voltage (resulting from the currents flowing in A and B directly from the transmitter) which establishes a current in C (long dashed arrow). The combined currents in C flow through R in a direction opposite to the current (solid arrow) which comes directly from the transmitter. Since these opposing currents are, by design, about equal, the receiver is not appreciably affected during transmitting, and sidetone is practically eliminated. The *transmitting volume efficiency* of both the anti-sidetone and sidetone circuits is about the same, but the *effective transmitting losses* for the anti-sidetone set are substantially less, principally because, when less sidetone is heard by the talker, he unconsciously raises his voice level until the sidetone heard is about equal to what he is accustomed to in ordinary conversation. Increased voice level results in transmission gain.

The range of line impedance conditions is usually much greater for the local battery anti-sidetone set which is used on the longer loops into outlying and rural areas. In order to provide an approximate balance for these line conditions, the local battery anti-sidetone circuit is designed with a balancing network that can be connected for a line impedance of 600, 900, or 1500 ohms (angle 50°).

Under average conditions, and depending upon the amount of sidetone reduction which may be effected in any particular case and upon the circuit design, a transmission improvement of about 5 db may be expected from use of the anti-sidetone in place of the sidetone circuit.

THE RINGING CIRCUIT of a common-battery subscriber set includes a ringer and condenser, but a condenser may or may not be required in this circuit for a magneto set. The common-battery sidetone circuit has a single condenser for use in both the receiver and ringer circuit, but with development of the anti-sidetone circuit a separate condenser is used in the receiver and ringer circuits. With the two- (split-) condenser arrangement and high-impedance ringers, the susceptiveness of the subscriber set to incoming line noise is materially reduced over the single-condenser set, especially on balanced party-lines (same number and type of ringers connected to ground on each side of the line). *Ringers* are manufactured in a large number of types (usually with permanent magnet yokes and biasing springs) and impedances for use with the several signaling systems and to meet signaling requirements. The types of *signaling systems* used include straight line (20 or 16 $^{2}/_{3}$ cycle), harmonic, superimposed, and pulsating ringing, of which *straight line ringing* is commonly used for individual line, two-party selective, four-party semi-selective, divided code, or non-selective bridged stations.

Harmonic ringers are designed for frequencies of 16.6, 20, 25, 30, 33.3, 42, 50, 54, 60, 66, and 66.6 cycles. These ringers have reeds of different weights which respond only to the frequency for which they are designed. The harmonic system provides selective signaling for up to five bridged or ten grounded ringing stations (five connected to ground on each side of the line).

Superimposed ringing requires the superimposing of d-c potentials (positive and negative) on the regular a-c ringing current in order to raise the peak positive and negative voltages sufficiently to break down a three-element, cold-cathode, vacuum-tube gap, which is connected from line to ground at each station. The tube may be so connected that its gap will break down when either the positive or negative combined potential is applied, but not both. The ringer is connected between one element of the tube and ground so that, when the tube functions, the a-c ringing current will operate the ringer. This type of ringing is used in four-party selective and eight-party semi-selective service. The control gap of one type of tube has a nominal breakdown of 70 volts, and the nominal main (ringing) gap sustaining voltage is 75 volts.

A-c relays, first used in this type of service, generally have been replaced by the tube.

Pulsating ringing employs positive and negative pulsations sent out over the line from the office. The station ringers are biased with a spring so that they operate when either the positive or negative pulsations are applied, but not both. This type of ringing may be used for four-party selective or eight-party semi-selective service.

Ringer impedances are designed to meet various service conditions. The low-impedance group will have d-c resistances of from 1000 to about 2500 ohms, and the high-impedance group of from 3500 to about 5700 ohms or more.

Usually each ringer has a biasing spring, one end of which is attached to the armature and the other end, by means of a winding cord, to an adjusting stud on the ringer frame. Thus, the armature can be tensioned or biased to the degree necessary for the type of service involved, and to avoid cross rings and also bell taps during switchhook and dial operation. Ringers may be of the polarized or non-polarized type, depending upon the type of ringing system employed.

The number of ringing bridges across a line or from either side of a line to ground is limited to a total capacitance of the ringer condensers not to exceed 2 μf where not more than eight such bridges are involved on non-polarized lines.

Condensers are usually of 2-μf capacitance when used in single-condenser sets and in the transmission circuit of two-condenser sets. The ringer circuit of these latter sets may have a condenser of 0.5- to 1.0-μf capacitance, depending upon the ringer impedance, or for some types of magneto stations the condenser may be omitted.

Hand generators of the three- or five-bar type, depending upon the length of line and number of bridged stations, are generally necessary in magneto subscriber sets for signaling the office or another subscriber on the same line. These generators carry a spring assembly which functions to open and close the generator circuit across the line when the generator is idle and operating, respectively. Figure 6 shows a view of a five-bar generator (Stromberg-Carlson make).

The switchhook spring assembly in subscriber sets generally consists of two pairs of make-break springs which automatically close and open the transmitter and receiver circuits when the handset is lifted from and restored to its cradle. These springs are designed to retain their adjustments over long periods of time. One manufacturer is using heavy-gage, phosphor-bronze with precious-metal contacts, which provide positive, nonmicrophonic contact.

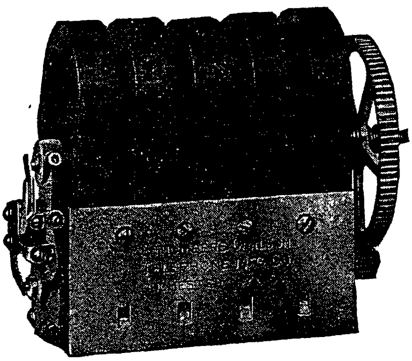


Fig. 6. Five-bar Hand Generator for Use in Magneto Subscriber Telephone Sets (Courtesy Stromberg-Carlson Co.)

Dials are required at all dial stations served by mechanical offices or private automatic exchanges of any type. The dial is the subscriber's only means at dial stations of securing connection with other dial subscribers or with toll or assistance operators in his exchange area. Figure 7 shows front and rear views of one type of dial assembly. The front view shows the finger wheel and stop and the number plate with its letter and number designations appearing through the finger wheel holes. The rear view shows the spring assembly, consisting of shunt (off normal) springs which short-circuit the transmitter and open the receiver circuits during dialing, and the impulse springs which open and close the subscriber line circuit, thus causing pulses of current to flow in this circuit at an average rate of about 10 pulses per second. The finger wheel, having been turned to the finger stop for any selected letter or number, is released, and as it returns to normal under spring action it is geared to operate the pulse pawl at a speed controlled by the governor. The pulse pawl alternately opens and permits the impulse springs to close as many times as there are units in the digit pulled.

Small filters consisting of inductance, capacitance, and resistance are generally bridged across the impulse spring contacts in order to reduce dialing interference to nearby radio receivers.

Housings of various types are provided, in which the required subscriber telephone set units described above are assembled. These housings are made of zinc alloy, plastics, steel for baseplates, and other materials, including rubber for cushioning. The various units are usually assembled on a universal baseplate, suitable for wall or desk-type sets.

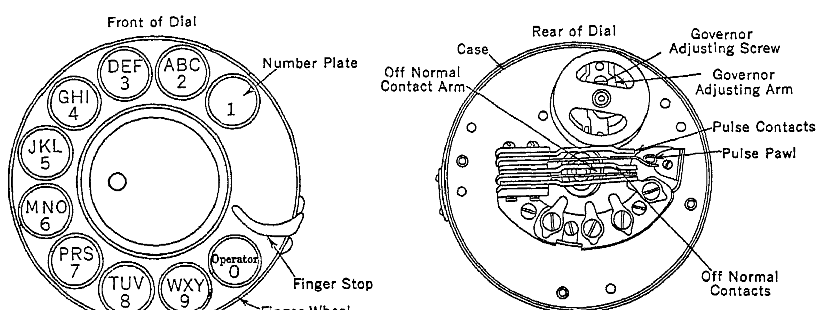


Fig. 7. Front and Rear View of Dial as Used with Dial Type Subscriber Telephone Set (Courtesy Bell System)

In one type of set, made by the Kellogg Switchboard and Supply Co., plug-in type induction coil (AST) and condenser units permit ready replacements to be made by pulling out the unit to be replaced and plugging in the new unit. The design of this company's set includes a universal permanently wired circuit, consisting of a stamped metal grid nested in the underside of an interconnecting block which mounts on a baseplate. This grid is connected through the block to screw terminals and pin jacks mounted on the upper side of this block. The induction coil has a three-way and the condenser a two-way switching unit to permit convenient circuit adjustments.

Some companies are now providing non-interfering or press-to-talk features in telephone sets intended primarily for use on multiparty rural lines, to permit listening on the line without the transmitter being cut into the circuit and without interfering with the ringing and dialing or with conversations in progress.

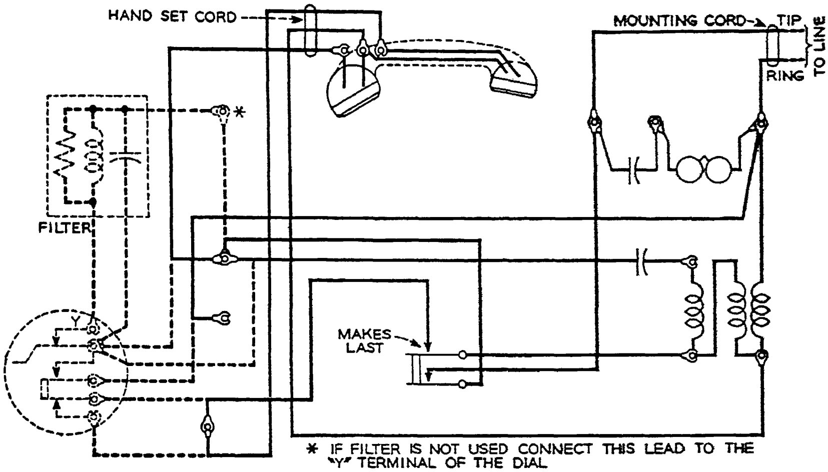


FIG. 8. Schematic of Wiring of Subscriber Hand Telephone Set—Common Battery Talking—Common Battery Signaling (Courtesy Bell System)

Figures 8 and 9 show, respectively, schematics of the wiring of a typical common battery talking-common battery signaling (CBT-CBS) and a typical local battery talking-common battery signaling (LBT-CBS) subscriber hand telephone set for individual, two-party selective, or four-party semi-selective service. The dial and filter circuits are required for dial service only.

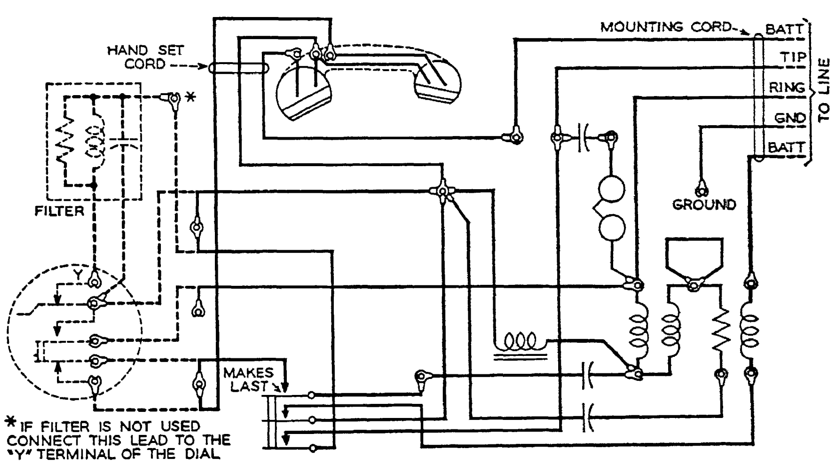


FIG. 9. Schematic of Wiring of Subscriber Hand Telephone Set—Local Battery Talking—Common Battery Signaling (Courtesy Bell System)

SUBSCRIBER SERVICES involving other than the regular subscriber telephone set just described require a wide variety of substation apparatus. Owing to limited space in this handbook, only brief mention can be made of some of the principal services currently rendered, which are:

1. *Station wiring plans*, whereby individual central office, PBX, or private lines may be connected at one or more telephone stations of which one or more or none may have a

regular attendant. In these cases, each station telephone may be equipped in its base with several pushbutton type keys, or separate key boxes may be provided which can be used for originating, answering, intercepting, holding, or transferring central-office calls on one or more lines, or connections can be established between stations without using either a central-office or PBX line. Various arrangements for controlling, locking out, grouping, or extending station calls are also available. Signaling between stations is usually by a pushbutton and buzzer circuit with battery or low voltage alternating current derived from the commercial power supply. Operating power may be supplied locally or from the central office.

2. *Loudspeaking and distant talking systems* may be employed between a master and one or more regular stations, facilitating communication, and such systems may be associated with regular telephone facilities and equipments. These systems require amplifiers, which are associated with each loudspeaking and distant talking telephone set, to provide adequate volume from the loudspeaker. Commercial power is usually employed for this equipment.

3. An *operator-type transmitter and receiver set*, with an associated key and jacks, may be provided at subscriber stations to facilitate handling messages over the telephone.

4. *Loudspeaker conference service* may be provided, where groups of people desire to listen to a local or distant talk and are assembled at one or more points. An amplifier and loudspeaker with a suitable subscriber set and switching keys are required for this service. Commercial power is used for this equipment.

5. *Coin collectors* of various types are located in public places where customers may place local or long-distances calls, either by prepayment or postpayment of the charges. Some of these collectors are placed in booths; others, in the open at locations convenient for public use.

6. *Subscriber amplifier deaf set equipment* provides for amplification of the incoming voice currents before reaching the subscriber's receiver. The incoming volume level is raised sufficiently, in many cases of deafness, for the subscriber to carry on a telephone conversation which would otherwise not be possible.

7. *Code calling systems* sound code signals at various points throughout a subscriber's establishment to notify certain employees or officials that they are wanted at the telephone or for some other reason. The controlling station on this system is usually located at the subscriber's PBX switchboard and is operated by the PBX attendant as a separate system, or the system may be actuated directly from a dial PBX or PAX system without an attendant.

This system requires code sending and station signal equipment, properly located and interconnected by wiring. Commercial a-c power supply is employed for operating it.

8. *Loudspeaker paging systems* are designed to provide a means of simultaneously transmitting messages or announcements verbally from a central location to a number of points within an establishment. These systems have a variety of uses from summoning a person to a telephone or directing employee activities, to providing information to a limited area.

9. *Subscriber telephone sets for explosive atmospheres* are available for use in mines, oil refineries, or munition plants. The equipment is designed to prevent sparking of the various parts under operation, causing explosions.

10. *Outdoor-type telephone sets* are provided for mounting in outdoor places for fire, police, taxicab, and other services. These sets are enclosed in cast-iron or wooden housings, which protect the equipment from weather conditions.

11. *Sound powered telephones* are provided in locations where it is desired to avoid possible central energy failures and where the distances between stations are relatively short, such as in an establishment, on a ship, or as portable field telephone equipment where batteries are not desirable. These telephones operate through action of sound waves striking the transmitter diaphragm and causing a variation in a magnetic field which produces electric currents in the telephone circuit of frequencies similar to those in the sound waves. The receiving apparatus functions quite like the regular telephone receiver.

12. *Program distributing systems* are designed to furnish program material, either from broadcasting stations or central program points, to hospitals, schools, hotels, business establishments, factories, homes, and many other locations. Program material may consist of music, speeches, announcements, and various other features of interest. This material is usually transmitted over wire lines from its source to one or more common amplifiers and thence distributed to subscribers over wire lines to loudspeakers at the various locations.

20. SUBSCRIBER STATION PROTECTION

Substation protection is required to protect subscribers and substation equipment from dangerous voltages and currents. The telephone plant is designed to withstand, with some margin, its normal operating currents and voltages. But there are other sources of electrical power, chiefly lightning and commercial power lines, which may under certain conditions impress large and destructive voltages, with resulting excessive currents, on the telephone plant, either by direct contact or by induction.

Danger from lightning in cities is less than in sparsely settled areas because of the shielding effect of buildings, trees, and various overhead structures. The danger from power-line contacts with aerial telephone plant is always present where the two types of plant are in close proximity, although material progress has been made in lessening this hazard over the years by improved methods of construction and protection. Stations connected to lines which are exposed to more than 250 volts between wires usually require protection and are classed as exposed stations.

The *maximum voltages* most commonly impressed on subscriber lines for telephone purposes range from 24 to 50 volts direct current and from 75 to 175 volts alternating current. Direct-current flow over the subscriber loop (individual line service) does not usually exceed 150 to 200 milliamperes and in most cases is less than about 100 milliamperes.

One type of *substation protector* widely used consists principally of two pairs of carbon protector blocks (open space cutouts) and two line fuses (one pair of blocks and one fuse for each side of the line), with associated spring and terminal holders, mounted on a porcelain base. Another design employs two metal plates, each of which connects with one side of the line and has a sawtooth inner edge positioned 0.004 in. from a grounded carbon block. The discharges take place across this gap. The ground electrode of the protector must be well grounded in every case.

CARBON PROTECTOR BLOCKS (sometimes designated as lightning arresters or dischargers) are manufactured in several designs, with the gap between the line and ground block varying from about 0.003 to 0.075 in. or more. The dielectric between the blocks may be air, or mica or acetate separators may be used. One design common throughout the United States consists of a grooved porcelain block with carbon insert for the line contact, and a solid flat (plain) carbon block for the ground contact. The line spring bears against the carbon insert set in the porcelain frame, and the porcelain frame bears against the ground block. The carbon insert is held in place by a glass cement of low melting point, and is forced against the ground block when arcing across the gap is sufficient to soften the cement.

The carbon insert is accurately positioned in its porcelain frame to provide a 0.003-in. gap for substation protection, or, when used at the junction of cable and open wire lines, this gap is 0.006 in. The peak breakdown voltages for these gaps are about 350 average and 550 maximum for the 0.003-in. gap and about 710 average and 1080 maximum for the 0.006-in. gap.

The fuses, being connected in series with the line and on the line side of the protector blocks, are designed to open each side of the line when the protector blocks discharge heavily or break down completely and when the resulting current through the fuses exceeds their ratings, which are usually 5 or 7 amperes, although fuses of other amperage may be used, depending upon operating company requirements.

Grounding of the protector requires a reasonably low-resistance ground (less than about 25 ohms), which may be obtained at the subscriber's premises by connecting to (1) the public water pipe system, (2) a private water pipe or well casing system, or (3) a driven ground (ground rod or pipe), preference being given in the order named.

A common ground with the secondary neutral of power distribution wiring is usually employed at the subscriber premises unless adequate separation between the telephone and power wiring on the premises can be maintained. This is to avoid excessive potentials being impressed on the telephone wiring, where separate grounds are employed, should there develop abnormal currents in the secondary neutral due to its becoming crossed with the primary power circuit.

Where extensive public water pipe or other systems having a low resistance to ground are available, common grounding should be used. Where such systems are not available, consideration must be given, in deciding on a separate or common ground, to the probability of high potentials being impressed on the telephone wiring in case of high currents in the secondary neutral.

Figure 10 shows one type of substation protector, widely used, with a fuse mounted along each side and the protector blocks held in a spring assembly set in a well in the center of the porcelain mounting block. The metal cap shown screws down over the blocks, excluding dirt and moisture.

Substations located at power stations generally require special protection. Whenever an abnormal condition on a power system results in ground current between the power station and an outside point on the system, the ground-potential rise at the power station above that of distant grounds will depend on the IZ drop in the power-station ground.

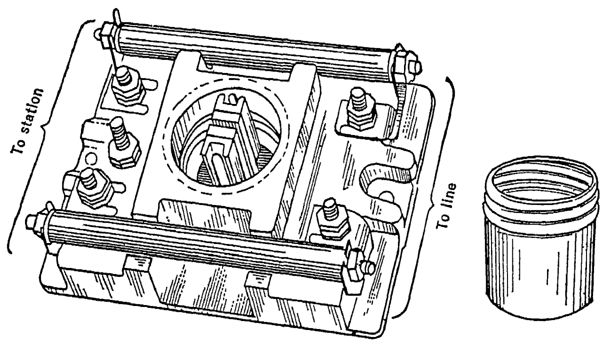


FIG. 10. Substation Protector (Courtesy Bell System)

Such a potential rise might exceed the breakdown potential of the usual telephone protector and render the telephone circuit inoperative at a time when it is most needed. Special protective devices, consisting of a neutralizing transformer or remote grounding of the telephone protector, may be considered in determining the type of special protection to use.

21. PRIVATE BRANCH EXCHANGE EQUIPMENT

Private branch exchange (PBX) equipment, as discussed here, includes manual and dial, cordless and cord, attended and unattended switchboards, whatever particular designations (such as PAX or PABX) are given to them by various companies. These boards range in size from the small ten-line system to the largest multiple type having a capacity of about 3000 lines or more.

MANUALLY OPERATED PBX BOARDS may be of the cordless type serving a few stations or of the cord type for the larger installations. These boards may also be operated in conjunction with PBX dial equipment.

A typical type of *manual cordless common-battery PBX board* has a capacity of twelve extensions (stations), five central-office trunks, and five connecting circuits. Each extension and each trunk terminates in a vertically mounted key unit which has three keys with levers in a vertical row. Each key lever can be operated either up or down or remain in its normal middle position. The upper and middle horizontal rows of keys, when operated up or down, connect their respective trunks or extensions to a common strapping between the keys in the board, and the lower horizontal row of keys does likewise when operated to the up position. In the down position, these latter keys bridge a holding coil across the trunk for the trunk keys and apply ringing current to the extensions for the extension keys, and the ringing keys are non-locking.

Each trunk has a visual drop and condenser bridged across it, and each extension has a visual signal in series, so that the central office and the extensions can each signal the attendant. A supervisory relay is provided in each connecting circuit, which circuit also provides talking battery for extension to extension connections, through a retard coil. For trunk to extension connections, talking battery is furnished over the trunk from the central office. The attendant's telephone set is connected to the first vertical key unit on the right side of the board.

Nominal power of 24 volts direct current is required for talking battery between extensions or between the attendant and the extensions and is usually furnished over cable pairs from the office, as is the required 20-cycle ringing current for the board. A hand generator is provided for the board for emergency use or where office ringing power is not available.

Incoming rings over a trunk operate the trunk drop, and the attendant answers by operating an idle connecting circuit key associated with the trunk, and also the attendant's corresponding key, both in the same horizontal row and to the same position, up or down. This connects the attendant's telephone to the trunk line.

If an extension is being called over a trunk, both trunk and extension keys are operated to corresponding positions and the extension ringing key is operated. When or before the extension answers, the attendant restores the attendant's telephone set key to normal, and when the supervisory signal indicates that the conversation is finished both the trunk and extension keys are operated to normal.

Extension to extension calls are established similarly by properly operating the calling and called extension keys.

Trunk calls can be held by the attendant operating the trunk holding key. The attendant or any extension may dial through the board to a dial office, if the telephone set is equipped with a dial. Outgoing trunk calls signal in to a manual common-battery office automatically.

Another type of *manual cord common battery PBX board* has a capacity of 320 extensions, 15 central office trunks, and 15 cord circuits. This board is of the two-panel, single-position, non-multiple type suitable for medium-size installations ranging from 80 to 320 extensions. Two boards may be operated side by side to increase the capacity. The trunks, central office or tie (to another PBX), and the extensions terminate on jacks between which the connections are established by means of the cord circuits. The trunks are ringdown incoming to and automatic signaling outgoing from the PBX. Each pair of cords has a talking and listening key and a ringing key for ringing on either the front or back cord, and double lamp supervision, except that supervision is provided on the extension cord only, for trunk calls. Each trunk and extension circuit is equipped with lamp signals for signaling the attendant, and up to 20 line relays may be provided for extensions to increase their signaling limit.

The attendant's telephone circuit contains a dial for dialing on dial trunks or extensions. The board may be operated in conjunction with dial PBX boards for intercepting calls or other services.

Talking battery is supplied for extension to extension connections through the single bridged retard coil in each cord circuit, but for trunk to extension connections talking battery is furnished from the central office. Nominal 24-volt battery required at the PBX for talking and circuit operations, and 20-cycle ringing power for ringing extensions, are usually provided over cable pairs from the central office. A hand generator is usually furnished for emergency use.

This board is designed to establish connections between local extensions and between these extensions and a manual or dial central office or other PBX.

Incoming calls from either a trunk or extension to an extension light an associated lamp signal at the board, and the attendant answers by operating the talking and listening key of the answering cord, which has been inserted in the calling jack. The connection is then completed by inserting the calling cord in the called extension jack and ringing this extension. Outgoing calls to the central office are handled over a trunk similarly, except that signaling the central office is usually automatic.

A type of *large manual cord common-battery multiple PBX board* has a capacity of 1520 extensions (without designation strips), 240 trunks, and 15 cord circuits (per position). The number of extensions may be increased by using a 34 $^1/_2$-in. in place of the usual 24 $^1/_8$ in. jack panel opening. There are two panels per position and four panels per multiple jack appearance for both extensions and trunks and one position per switchboard section.

The trunks and extensions terminate on series cutoff type jacks in the face of the board, and each jack has a multiple line lamp (for manual operation) associated with it. Since each extension or trunk appears in the jack multiple at every fourth panel throughout the board, the multiple lamps can be arranged to light, for an incoming call, at each appearance up to a total of four, thus attracting the attention of a greater number of attendants and improving answering performance. Line relays may be provided for the extensions to increase their signaling range. The manual trunks are usually ringdown incoming and automatic outgoing, and for a dial office they are of the dialing type.

The cord circuits are of the bridged-impedance series-condenser type, which provide battery feed to each cord of the pair separately through bridged retard coils. The tip and ring of the cord pair each has a condenser in series which prevents d-c flow in one half of the cord circuit affecting d-c flow in the other half. Talking battery is thus supplied each extension from the PBX cord circuit for extension to extension calls and to the extension for tie trunk to extension calls. For central-office trunk to extension calls the cord circuit is so arranged that extension talking battery is furnished from the central office, except where PBX long line circuits are provided at the PBX.

Each cord pair has a talking and dial key and a ringing key for ringing on either the front or back cord. Double lamp supervision is provided for each cord circuit on extension to extension calls, and single supervision for trunk to extension calls. A position dial may be cut in on any cord for dialing by operating the talk and dial key.

The distributing frames are enclosed in sections at the head of the switchboard. Facsimile sections may also be located in the switchboard line-up. The board may be operated in conjunction with dial PBX boards for intercepting or other services.

The power required for operation of this board is a nominal 48-volt d-c source for talking and circuit operations, and 20-cycle ringing current. Owing to the usually relatively heavy battery drains, storage batteries are provided at the PBX and charged by a separate power unit at the PBX or, where economical and practicable, over cable pairs from the central office. Ringing current is also generally supplied to the PBX over cable pairs from the office, and each position is equipped with a hand generator for emergency use.

This board is designed for establishments, such as large department stores, institutions, and organizations, where PBXs of the smaller types are not adequate. It may be operated in manual or dial office areas and in conjunction with other manual or dial PBXs.

Operation is similar to a regular exchange manual switchboard except for the several special features that may be provided to meet the individual subscriber's needs.

A typical *medium-size dial PBX unit* of the step-by-step (SXS) type has a capacity of up to 79 extensions, 10 central office and 10 attendant's trunks, for two-digit dialing. The SXS equipment consists principally of 20 line finders, 100 line and cutoff relays, and miscellaneous equipment mounted on the side of the switch frame, and 20 selector-connectors and 19 trunk equipments mounted on the reverse of the frame.

This PBX unit functions with a suitable companion manual PBX containing a jack appearance for each trunk and extension for the purpose of receiving, originating, and intercepting calls that cannot be handled by the SXS unit. Incoming calls reach the manual attendant, who dials the called extension over an attendant's trunk. Outgoing calls may be handled and calls requiring special treatment are usually handled at the manual board. The SXS unit handles direct calls between extensions and from extensions to the central office, associated manual PBX board, and other PBX boards equipped to receive dialing pulses.

Power requirements for the SXS unit consist of nominal 48 volts direct current for talking and circuit operations, and machine ringing current. D-c power is usually supplied by a small power unit located at the PBX, or over central-office cable pairs if the load permits and pairs are available. Ringing current is generally furnished over central-office cable pairs.

Dialed calls are completed by the usual step-by-step processes (described elsewhere in this section). Calls handled at the manual PBX may either be dialed in dial office areas or completed over automatic signaling trunks in manual office areas.

Dial PBX units of the step-by-step type having capacities up to 3200 extensions, depending upon the number of extensions required, are available. Two-digit dialing (under 100 extensions) requires line finders and selector-connectors (operating as a selector on the upper levels and a connector on the lower levels) while four-digit dialing requires first and second selectors and connectors. If incoming dial repeating trunks are provided, they terminate on incoming connectors for two-digit and on incoming selectors for three- or four-digit dialing. The line finders have a 200-point bank; the other switches have 100-point banks.

These dial PBX units are usually associated with manual PBX units, the latter receiving, originating, and intercepting calls that the SXS unit is not arranged to handle. The SXS unit handles calls between extensions and calls to the central office or to another PBX which are capable of receiving dial pulses from the extensions or over dial repeating trunks. The extensions of the SXS unit may be reached direct from other PBXs if proper dialing facilities are provided, but calls to these extensions from the central office are usually handled at the manual PBX board.

PROTECTION for PBX trunks and extensions at the PBX is the same as for regular subscriber to central-office lines of the same classification, that is, exposed or non-exposed. In addition, owing to paths to ground through the equipment in the PBX board, which may permit foreign currents large enough to cause trouble to flow through the equipment, heat coil type fuses having the same operating characteristics as central-office heat coils (except that they open rather than ground the line wires) are provided in each exposed trunk and extension at the PBX. In the case of large PBXs a section of fine-gage cable sometimes is used in place of the 7-ampere line fuses. This is similar to the fusing provided at central offices (see article 6).

ORDER TURRETS, manufactured in various types and capacities, are employed in business or service establishments, such as department stores, telegraph offices, taxicab companies, and newspaper offices, to receive orders for goods or requests for services.

One type of order turret for multidepartment businesses has a capacity of one two-way trunk and one outgoing trunk from the turret to the subscriber's PBX and one incoming trunk from the PBX to the turret. The incoming customer calls are received at the PBX board and routed over trunks to the proper department turret, where an attendant receives the customer's request. The turret serves as a small switching unit between a single telephone in the departments thus equipped and the PBX, and it avoids the installation of several telephones at each location.

An order turret system, where incoming calls are concentrated in one line-up of turret positions, is also available. Incoming customer calls from either a manual or dial central office or PBX are received in incoming trunk circuits, allotted by an allotter circuit to a sequence storing circuit, and thence released to an idle attendant in the sequence in which the calls are received. Thus the calls are distributed automatically as rapidly as turret attendants become idle and with minimum loss of time in handling. This type of system has a capacity of 120 incoming trunks (with first group of selectors), 110 attendant's trunk circuits, and 110 attendant's positions. A small order turret located in front of each attendant has four keys and two lamps, a trunk, and a calling waiting signal for receiving and originating calls. Calls can be held, released, or transferred by means of these keys. Usually a manual or combined manual and dial PBX board is associated with this turret system. Power supply is 20–28 volts direct current, for which a local automatic power plant is provided. Ringing current is usually furnished over cable pairs from the central office.

SECTION 18

TELEGRAPHY

TELEGRAPHY

THEORY

By John D. Taylor

Fundamentally, the process of *communicating* by telegraph consists of sending electrical impulses by wire or radio from a transmitting to a receiving point. These impulses are so selected, arranged, and transmitted (in sequence), that they are received, interpreted, and recorded as intelligible characters at the receiving point. Such interpretation may be made by the ear listening to the click of a telegraph sounder or the tone in a head receiver or loudspeaker, or by automatic and mechanical devices.

Basically, it was necessary to develop *codes* (intelligence-bearing arrangements of electrical impulses) and to provide transmitting and receiving devices and interconnecting channels, capable of satisfactorily passing these codes between the desired locations.

1. METHODS OF TRANSMISSION

Telegraph codes or signals are transmitted electrically (1) by open-wire land lines or cables or short submarine cables, using either direct or alternating current, (2) by long submarine cables (oceanic), using only direct current, and (3) by radio, using only alternating current.

Direct-current telegraphy employs a d-c source of power, which is either interrupted, reversed in polarity, or altered in magnitude by the transmitting mechanism. The line current frequency with present high-speed telegraph systems does not usually exceed about 200 cycles per second, and thus systems using this type of transmission are generally referred to as *low-frequency systems.*

Alternating-current telegraphy employs, for each transmitting channel, a relatively narrow band (about 170 to 300 cycles wide) of a-c frequencies, which are modulated by the telegraph signals, being transmitted, before being applied to the line. A basic block of frequencies, approximately 3000 cycles wide, is used to provide a group of carrier channels, the number of channels per block depending on the width allocated to the individual channels. Carrier telegraph systems utilize one, two, four, or more basic blocks of frequencies in each direction of transmission, the blocks being stacked one above the other in the assigned frequency range by translation or second-modulation, as discussed in article 3, Section 17.

The transmission and reception of signals may be either manual or automatic. In *manual operation,* the transmitted signals are formed by a hand-operated switch or *sending key,* and the signals are received either on an electromagnetically operated sounder, which converts the signals into audible clicks, or on a Morse recorder, which makes an ink record of the received impulses. These audible or recorded signals are interpreted by an operator, who typewrites the message manually. *Automatic operation* provides for the formation and transmission of the signals by an automatic circuit-interrupting device or *transmitter,* which may be controlled either directly or indirectly by a group of keys similar to a typewriter keyboard. The received signals are automatically interpreted and recorded in typewritten form by a *printer.* In some cases, notably in the operation of submarine cables, a combination of automatic transmission and manual reception is employed. The transmitting and receiving equipment, and the associated line terminal apparatus, are so closely related that the entire assembly is usually referred to as a *telegraph system.*

2. CODES

Characters of the alphabet, figures, and punctuation marks are transmitted as combinations of marking and spacing impulses. A *marking signal* or impulse is one which causes the receiving apparatus to be operated, and a *spacing signal,* which separates successive marking signals, places the receiving apparatus in the unoperated condition. Spacing signals may be intervals of no current, impulses of opposite polarity to the marking signals, or impulses of different current value from the marking signals. *Two-element codes,* employing only two conditions of current, positive and negative, or current and no current, are mainly used on land line circuits. *Three-element codes,* which use both positive

Telegraph Codes

	Morse	Continental	Cable Morse
A	· —	· —	+ − 0
B	— · ·	— · · ·	− + + + 0
C	· ·	— · — ·	− + − + 0
D	— · ·	— · ·	− + + 0
E	·	·	+ 0
F	· — ·	· · — ·	+ + − + 0
G	— — ·	— — ·	− − + 0
H	· · ·	· · · ·	+ + + + 0
I	· ·	· ·	+ + 0
J	— · — ·	· — — —	+ − − − 0
K	— · —	— · —	− + − 0
L	——	· — · ·	+ − + + 0
M	— —	— —	− − 0
N	— ·	— ·	− + 0
O	· ·	— — —	− − − 0
P	· · · · ·	· — — ·	+ − − + 0
Q	· · — ·	— — · —	− − + − 0
R	· · ·	· — ·	+ − + 0
S	· · ·	· · ·	+ + + 0
T	—	—	− 0
U	· · —	· · —	+ + − 0
V	· · · —	· · · —	+ + + − 0
W	· — —	· — —	+ − − 0
X	· — · ·	— · · —	− + + − 0
Y	· · · ·	— · — —	− + − − 0
Z	· · · ·	— — · ·	− − + + 0
1	· — — ·	· — — — —	+ − − − − 0
2	· · — · ·	· · — — —	+ + − − − 0
3	· · · — ·	· · · — —	+ + + − − 0
4	· · · · —	· · · · —	+ + + + − 0
5	— — —	· · · · ·	+ + + + + 0
6	· · · · · ·	— · · · ·	− + + + + 0
7	— — · ·	— — · · ·	− − + + + 0
8	— · · · ·	— — — · ·	− − − + + 0
9	— · · —	— — — — ·	− − − − + 0
0	——	— — — — —	− − − − − 0
Period	· · — — · ·	· · · · · ·	+ + + + + +
Comma	· — · —	· — · — · —	+ − + − + −
Semicolon	· · · · ·	— · — · — ·	− + − + − +
Colon	— · — · ·	— — — · · ·	− − − + + +
Interrogation	— · · — ·	· · — — · ·	+ + − − + +
Quotation	· · — · — ·	· — · · — ·	+ − + + − +

Short figures used in Continental and Cable Morse Codes where no confusion would result:

1....	· —	6....	— · · · ·
2....	· · —	7....	— — · · ·
3....	· · · — —	8....	— · ·
4....	· · · · —	9....	— ·
5....	·	0....	—

Fig. 1. Telegraph Codes—Morse, Continental, and Cable Morse

and negative polarities and also a zero or no-current interval to separate groups of impulses, are employed chiefly on long non-loaded submarine cables, where the comparative freedom from extraneous interference removes the chief objection to the use of a zero interval for the separation of signal groups.

The Morse Code is used almost exclusively in the United States on hand-worked land lines; the *Continental Code* has been adopted by almost all foreign telegraph administrations, and is universally employed on radio telegraph circuits. In these two codes, the dot signal is the basis of time measurement. A dash is three times the length of a dot; the spaces separating successive impulses in a combination are of dot length; the spaces between impulse groups are equal to three dots; and a space equivalent to six dots is used to separate words. The Cable Morse Code is used almost entirely in the operation of long submarine cables. All impulses and spaces are of equal length, a positive impulse representing a dot and a negative impulse representing a dash.

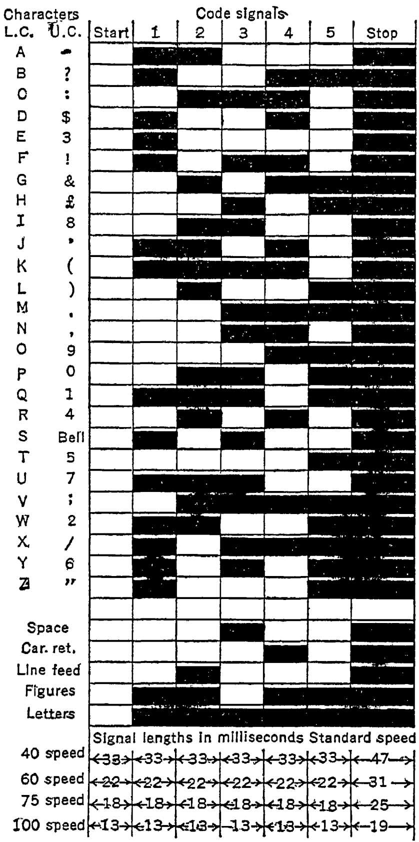


FIG. 2. Teletypewriter Code (Courtesy Bell System)

In calculating speeds, the average Morse character is equivalent in length to approximately 8.5 impulses of dot length, or about 4.25 cycles, while Continental characters average about 9 impulses or 4.5 cycles each. The Cable Code averages 3.7 impulses or 1.85 cycles per letter. These figures take into account the frequency with which the various letters of the alphabet occur in ordinary telegraph traffic.

Figure 1 shows the combinations of dot and dashes used in the Morse and Continental Codes, and the positive and negative polarity and zero current intervals used in the Cable Morse Code.

While the Morse and Continental Codes are designed for *manual* operation, the Cable Morse Code is transmitted *automatically*. The automatic transmission of telegraph signals is almost universally employed for land line, cable, and radio telegraph systems, because of its greater speed and ease of operation, resulting in large economies. Manual operation, when and where used, is, in general, confined to short-haul traffic and special services of such a nature as not to warrant or be adaptable to automatic operation.

In developing automatic telegraph systems, it was necessary to *design machines or devices*, which would automatically send and receive electrical impulses (positive or negative) and to *devise individual arrangements of these impulses*, representing intelligible characters. It was further required that these arrangements could be set up manually for transmission, and sent, received, and interpreted automatically at speeds suitable to the transmitting medium.

One type of sending and receiving device for use in automatic systems is the *teletypewriter* or *teleprinter* (described in detail in article 7 of this section), which utilizes a *5-impulse code* for each character with a single impulse for starting and also for stopping both sending and receiving units between characters. The code devised for the teletypewriter, which is practically identical with that of the teleprinter, is shown in Fig. 2, in which L.C. and U.C. are abbreviations for lower- and upper-case characters on the typing keyboard and the black and white spaces are the marking and spacing impulses, respectively. The signal impulse lengths in milliseconds (ms) are also shown for different operating speeds (words transmitted per minute).

Of the 32 separate combinations available, using 5 equal-length impulses for each character, the combination of all spacing impulses has no character assigned to it and is not transmitted.

The line frequency for the teletypewriter operation depends upon the number of words per minute being transmitted. For 60-speed transmission (about 60 words per minute) the shortest signal element is about 0.022 sec (see Fig. 2), and the line frequency is

$$1/(0.022 \times 2) = 22.7 \text{ cycles per second}$$

3. TELEGRAPH SIGNALS

Electrical impulses must be formed by the transmitting device, transmitted through a connecting medium, and finally received and interpreted by the receiving device, so that the received message is identical with the original sent message and is efficiently transmitted. However, owing to the electrical characteristics of telegraph circuits and associated apparatus, telegraph signal currents are generally more or less modified (electrically) in transmission, and, if suitable corrections were not made, these modifications or distortion would, in many cases, cause errors in the received message.

Telegraph signals, in d-c operation, are classed as (1) *neutral* (current flows over the line in either direction for the operated or *marking* position and no current flows for the non-operated or *spacing* position of the line relays) or (2) *polar* (current flows over the line in one direction for the *marking* and in the opposite direction for the *spacing* position of the line relays). For either type of signal, the change of current from mark to space or space to mark is known as a *transition.*

4. WAVE SHAPES

NEUTRAL SYSTEM. The change in line current values, with respect to time, may be plotted to show the *wave shape* of the telegraph signal for any telegraph circuit, as shown in Fig. 3 for a neutral circuit transmitting the Morse signal (dot, dash) representing the letter *A*. For such a circuit, *A* in Fig. 3 shows the wave shape of the current, if the times of building up and decaying to the steady-state values are neglected, *B* shows the effect on the wave shape of series inductance (line relay or composite set winding), and *C* shows, by the shaded area, the additional effect on the wave shape of the composite set or other condensers.

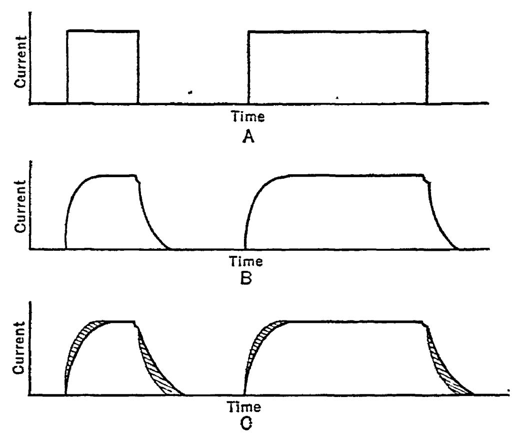


FIG. 3. Wave Shape for Letter A—Neutral Telegraph System (Courtesy Bell System)

With reference to *C*, Fig. 3, when the telegraph sending key closes, the inductance in the circuit opposes any sudden change in current value, and part of the current is diverted from the line to charge the composite set condensers to about the applied battery potential, as the line current builds up to its maximum value; when the key opens, the current immediately starts to decay and, although partly sustained by flow of current from the condensers and the opposition of the inductance to current change, shortly reaches its minimum value.

Current wave shapes are important factors, which affect telegraph relay performance and adjustments. For any given relay, there is a definite operating and release (less than the operating) current value for given operating conditions. Figure 4 shows a schematic d-c circuit, containing a neutral type of telegraph relay, milliammeter, battery, and rheostat. The retractive spring tension is controlled by adjusting screw S_1, the armature to pole piece air gap by S_2, and the armature travel distance by S_3 and S_4.

Figure 5 shows the effect of relay adjustments on operating and release current values (indicated by the black dots, O, R, O_1, and R_1 on the curves), and on the effective length (T and T_1) of the telegraph signal. In telegraph parlance, the shorter marking signals are called "light" and the longer marking signals "heavy." In practice, relay adjustments are made, as required, to provide satisfactory received signals from usually distorted wave shapes, caused by the electrical or mechanical characteristics of the line and associated equipment or changes therein.

Effective signal length might also be increased or decreased for a given neutral telegraph circuit with a fixed relay adjustment by raising or lowering, respectively, the applied circuit voltage. However, since line current values are limited, in practice, by crossfire into other telegraph circuits or by interference with telephone transmission, this method of controlling the signal length is restricted and not usually employed.

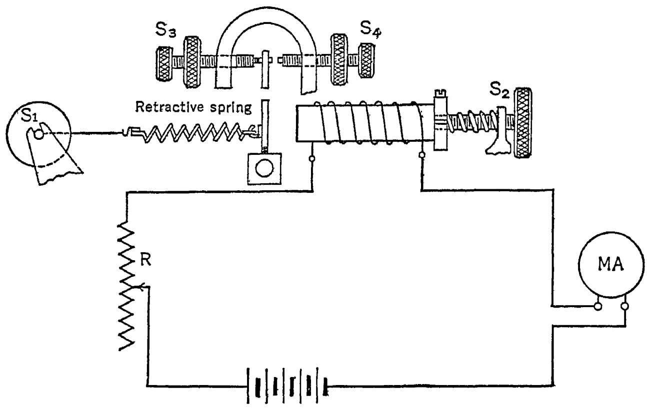


FIG. 4. Schematic of a D-c Telegraph Circuit with Neutral Relay (Courtesy Bell System)

Relay adjustments are also affected by line conditions, such as leakage of line currents to ground through tree or other contacts or the direct capacitance between the line wire and ground. In general, for open-wire telegraph circuits this leakage factor varies as among

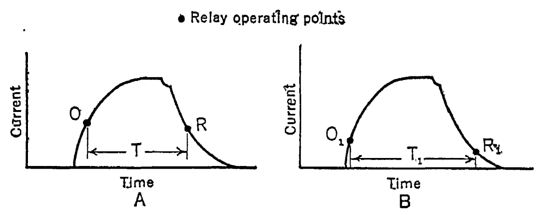


FIG. 5. Effect of Relay Adjustment on Telegraph Relay Operation (Courtesy Bell System)

wires and increases materially from the dry to wet weather condition. Figure 6*a* shows a neutral telegraph circuit with grounded battery at one station and Fig. 6*b* the same circuit with grounded battery at both stations, and for both circuits the leakage to ground,

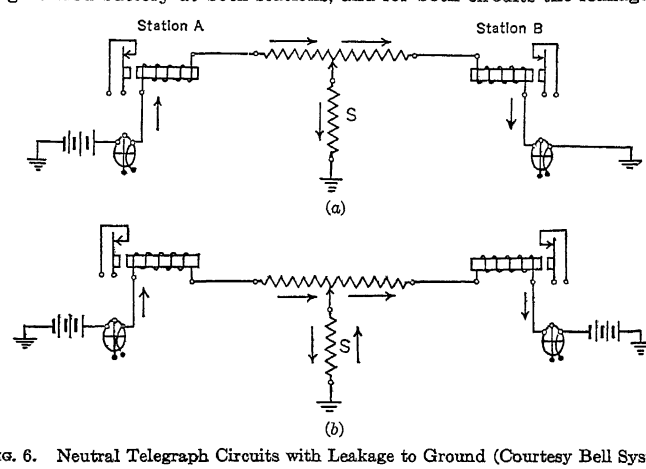


FIG. 6. Neutral Telegraph Circuits with Leakage to Ground (Courtesy Bell System)

which may be concentrated principally at one or more points or generally distributed along the line, is represented by the resistance S.

Figure 7 shows leakage effects on the current curve for the circuit (Fig. 6*a*). The curve in Fig. 7, *A*, represents the ideal condition of no leakage. The curve in *B* represents the leakage current alone through *S*. Since the inductance through *S* is less than through station B, the leakage curve in *B* is steeper than the curve in *A*, and this resultant, which affects station A relay, may be about as shown in *C*. In this curve, the effective signal length T_c is greater (heavier signal) than T_a in *A*. The leak *S* also shunts the path through station B and tends to decrease the line current to this station, as well as to flatten the current curve, as shown by *D*.

The wide variation of effective signal lengths, T_c and T_d, through station A and B relays, in series, would not occur if half of the total line battery voltage was applied at each station, oppositely poled to the line, as shown in Fig. 6*b*. Since each battery supplies current to the leakage path *S*, but in opposite directions, the resultant current through this path will generally be small and the line current through the station A and B relays will be more nearly equal and stabilized than for single end battery feed, thus improving signal transmission.

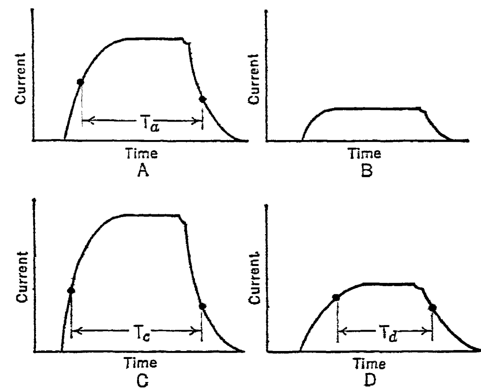


FIG. 7. Effects of Leakage on Current Curves for the Neutral Telegraph Circuit of Fig. 6*a* (Courtesy Bell System)

Telegraph current wave shapes formed by direct current are, in effect, somewhat similar to a-c wave shapes and contain various harmonics of the fundamental frequency, as will be discussed later. Owing to the normal line wire capacitance to ground (irrespective of other leakage paths), the a-c components of the telegraph currents will be shunted, in some degree. Thus, the line capacitance not only reduces the effective current at the receiving station but also tends to distort the current wave shape and to limit the length of operating line section between repeater points.

As indicated in Fig. 5, the relay operating points occur on a current curve between points of transition. The change from the spacing to the marking condition is designated *space-to-mark (S-M) transition*, and the change from the marking to the spacing condition is designated *mark-to-space (M-S) transition*. At the sending end of a telegraph circuit, the S-M transition takes place when the key is closed, and the M-S transition occurs when the key is opened.

There is a certain *time delay* from the closing of the sending key to the operation of the receiving relay and from the opening of the key to the release of the relay. In the first case, this delay is designated *space-to-mark transition delay (S-MTD)* and in the second case, this delay is designated *mark-to-space transition delay (M-STD)*, as shown in Fig. 8*b*.

In Fig. 8*a* the line capacitance is represented by the dotted condenser path to ground. When the sending key is closed, this capacitance is charged by current flowing over the line from the sending station, and the current is retarded in building up to its full value by the inductance in the circuit, as indicated by the sloping wave shape in Fig. 8*b*. The current which flows continuously in the biasing winding of the receiving relay produces a constant magnetic field tending to hold, in this particular circuit, the receiving relay armature against its spacing contact. When a marking signal current is received, the stronger opposing magnetic field set up by the larger marking current in the operating winding of the receiving relay causes the relay armature to move to its marking contact.

Although the operating points of a relay depend upon its design and adjustments, it may be assumed, for the purposes of this discussion, that the relays in Fig. 8*a* will operate on 33 and release on 27 ma of direct current, or plus and minus 3 ma from the 30 ma of biasing current, as shown in Fig. 8*b*. When the operating current increases to 30 ma, the effective operating current in the receiving relay is then zero, but when it reaches 33 ma, operation occurs and continues until the line current, decaying on open circuit, decreases to 27 ma, when the relay releases.

The S-MTD period is the time it takes for the operating current to increase from zero to 33 ma, and the M-STD period is the time it takes for the current to decay from its maximum value of 60 ma to the release value of 27 ma. These periods vary from a fraction of a millisecond (ms) to several milliseconds, being determined, for a given circuit and adjustments, solely by the circuit characteristics. For a given circuit and operating conditions, each transition delay (S-MTD) will be the same for repeated signals, and the same holds true for each delay (M-STD), but the S-MTD delays may not be equal to the M-STD delays.

Each mark, of whatever length, begins with an S-M transition and ends with an M-S transition. The S-MTD period reduces and the M-STD period increases the effective

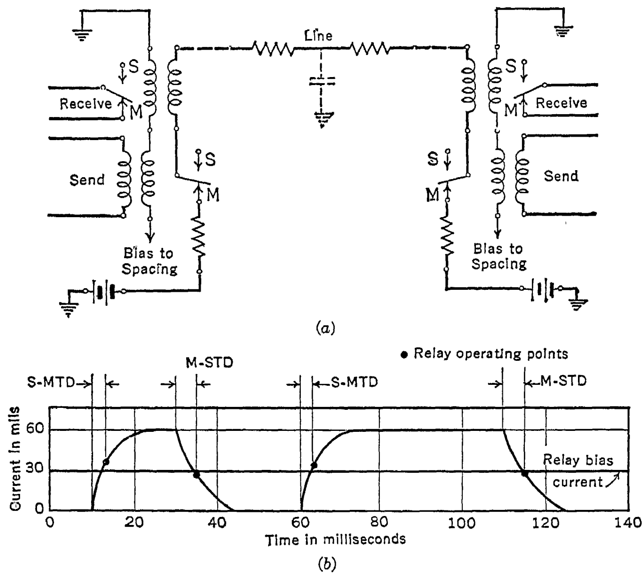


FIG. 8. (a) Neutral Telegraph Circuit and (b) Signal Wave Shape (Courtesy Bell System)

length of the mark, so that, if these two delay periods are equal, the length of the mark will not be changed by transmission over the circuit. Each space, of whatever length, begins with an M-S transition and ends with an S-M transition. The M-STD period reduces and the S-MTD period increases the length of the space, so that, if these two delay periods are equal, the length of the space will not be changed by transmission over the circuit. The transmission of signals is considered perfect if the received effective marks and spaces are exactly the same length as the sent marks and spaces.

POLAR OPERATION. Wave shapes in polar telegraph systems are affected by circuit inductance, capacitance, and leakage somewhat as in neutral telegraph systems. Figure 9

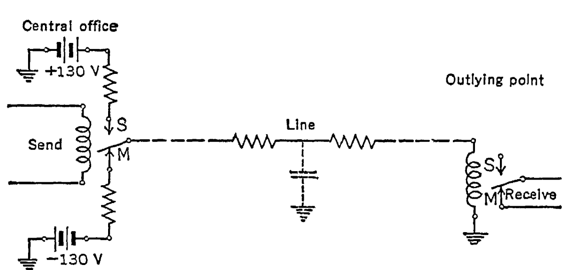


FIG. 9. Simplified One-way Polar Telegraph Circuit (Courtesy Bell System)

shows a simple one-way polar telegraph circuit, arranged to send −130 and +130 volt impulses from the central office (sending point) to an outlying receiving point, having series line resistance and capacitance between the line and ground.

Assuming that the sending-end connections are adjusted to provide normally steady-state line currents of +35 ma (marking) and −35 ma (spacing), as shown in Fig. 10, the line capacitance to ground will delay the change of line current from spacing to marking (S-M transition) and from marking to spacing (M-S transition).

The M-S and S-M wave shapes (Fig. 10) are identical in form and symmetrically located about the zero line, and the S-MTD and M-STD are equal. Thus, since the sending end potentials are equal and of opposite sign, since the circuit resistance remains constant for both positions of the sending relay armature, and since the operating points of the receiving relay are symmetrically located at about the middle of the wave shapes, the received polar signals are unbiased.

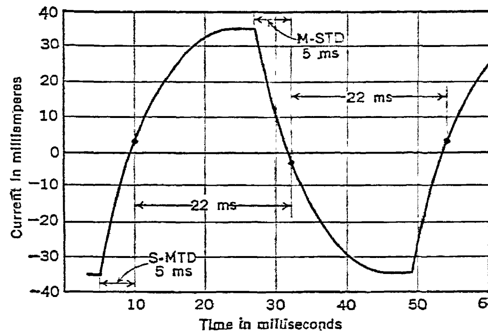


FIG. 10. Signal Wave Shape in Typical Polar Telegraph Circuit (Courtesy Bell System)

TELETYPEWRITER AND TELEPRINTER OPERATION. In teletypewriter or teleprinter operation, there are five equal length impulses, each of which may be negative (mark) or positive (space), in accordance with the code of signals (Fig. 2) for this type of operation. These five impulses represent the character to be transmitted. Two additional impulses are sent with each character, one starting and one stopping the machines.

If a series of impulses, representing, for example, the letter *D*, are plotted with time as the horizontal axis, as shown in Fig. 11, it will be noted that the wave shape, in the upper

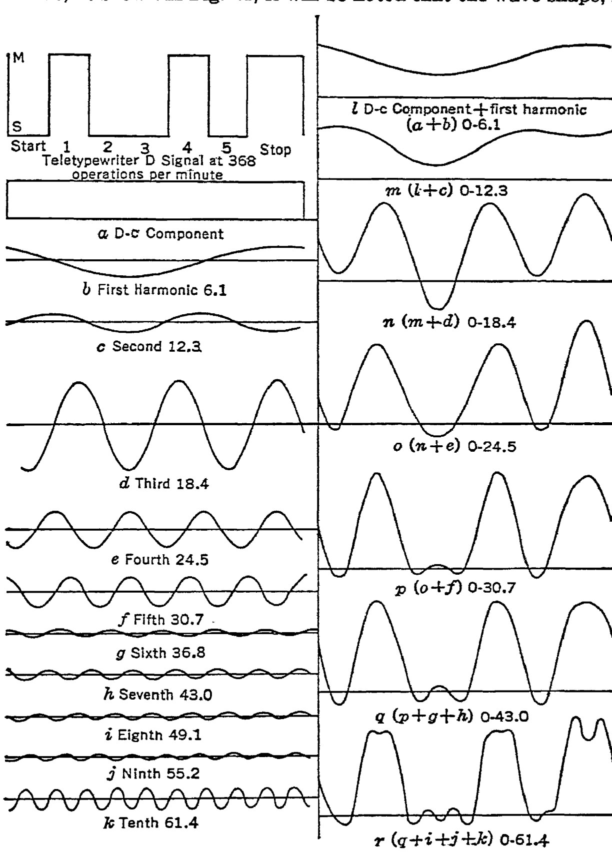


FIG. 11. Analysis of Wave-shape Components of Teletypewriter Character D (Courtesy Bell System)

left-hand corner of the figure, which is impressed on the line at the transmitting end of the circuit, has "square" corners. However, if such a wave is analyzed, it will be found to contain an infinite number of component frequencies, all of which, in practice, cannot and need not be transmitted to the receiving terminal.

Though a very simple receiving device might be used, if all the components could be faithfully transmitted over a circuit, such a circuit could not be economically provided. Also, with an ideal receiving device, it would only be necessary to transmit a maximum frequency equivalent to that which considers the duration of each unit signal element of the code as $^1/_2$ cycle (1 cycle would be the time involved from the end of element 1 to the end of element 3 of the square wave in the upper left-hand corner of Fig. 11). This condition is closely approached for long submarine cable telegraph systems, where the high conductor cost warrants expensive terminal equipment for reasons of efficiency and economy. However, for land line teletypewriter commercial use, neither expensive circuits nor elaborate terminal arrangements are justified, and the facilities and equipment that are provided represent a compromise on a cost balance basis.

For good telegraph transmission with comparatively simple equipment, it is desirable that the received signal contain a substantial part of the second and third harmonics of the frequency of the unit signal element. For 60-speed teletypewriter signals, the shortest signal element is 0.022 sec, which is equivalent to $^1/_2$ cycle. The fundamental frequency at this speed then would be $1/(0.022 \times 2) = 22.73$ cycles per second, the third harmonic of which would be about 68.2 cycles per second.

If one teletypewriter character is continuously repeated at 60-speed, the signal wave (there being one per character) repeats itself about 6.1 times per second, so that, for the purpose of discussing the signal components in teletypewriter operation, the wave may be considered as composed of a d-c component and harmonics of a fundamental frequency of 6.1 cycles, rather than harmonics of the signal element frequency. Thus, in Fig. 11, the left-hand column shows several of the more important harmonic components of the D signal, their relative magnitudes and phase relationships, the first harmonic (curve b) being shown as a sine wave of the same time period as the *overall signal;* the right-hand column shows the combined resultant of the d-c component and the overall signal harmonics, added successively, as indicated.

Since the fundamental signal element frequency for 60-speed operation was shown previously to be 22.7 cycles per second, theoretically the character D could be interpreted correctly by an ideal receiving device if components, in correct phase relation, up to and including the fourth harmonic of the overall signal (curve o in right-hand column) were received.

In practice, overall signal harmonics up to about the tenth (corresponding to about the third harmonic of the signal element frequency) are transmitted, which gives a signal wave similar to curve r and which resembles somewhat the square-corner wave in Fig. 11.

CARRIER TELEGRAPH. Carrier telegraph systems employ modulated alternating currents of different frequency bands. The voice-frequency (low-frequency) system utilizes the band of 255 to 3145 cycles for 18 channels of 170 cycles per channel. The modulated output of this system may be impressed, if desired, on broad-band telephone carrier or radio channels of higher-frequency bands, making use of such channels for telegraph rather than for telephone service. The high-frequency system for telegraph was the first one developed. It is for open-wire facilities and operates in three different frequency assignments, all of which lie within the overall range of 3.33 kc to 11.25 kc. These systems will be described later in this section.

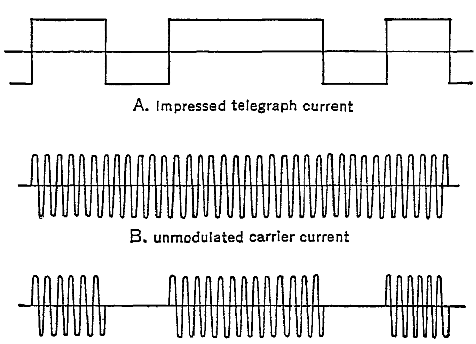


FIG. 12. D-c Modulation of Carrier Currents—Telegraph Carrier Systems (Courtesy Bell System)

As previously described, the basic signal element is a square wave produced by making and breaking a d-c sending circuit or by applying opposite potentials (one at a time) to the transmitting device. Generally, the carrier wave is supplied continuously to the transmitting device and is interrupted (modulated) in accordance with the d-c signals being produced by the local sending circuit and impressed on the carrier wave at this device. Figure 12 shows the impressed telegraph current (polar operation), the unmodulated carrier current, and the modulated carrier wave. The received segments of carrier current signal are rectified in the demodulator circuit, and the resulting unidirectional plate current pulses act on a receiving d-c relay. These received d-c current signals, being similar in pattern to the sent d-c signals, actuate the receiving relay in accordance with the sent signal.

Since the sending d-c signal contains, for practical reasons, up to about the third harmonic, which at 60-speed operation is about 70 cycles, both carrier sidebands transmitted will also contain this band width, or, for both bands at this speed, the frequency spread would be about 140 cycles. For higher or lower speeds, this spread would be greater or less.

5. DISTORTION

An ideal (perfect) telegraph circuit reproduces telegraph signals at the receiving end exactly as they are impressed at the sending end, with respect to length, but not necessarily amplitude, of the component marks and spaces. The time of travel of the signal over a circuit is usually not important, even though such time exceeds the duration of a unit signal element.

Distortion in telegraph transmission is thus determined by comparing the length and relative arrangement of the signal elements as sent with the length and relative arrangement of such elements as finally delivered by the receiving device.

The overall or total resultant distortion of signals, for a given telegraph circuit, consists of two principal types of distortion, namely systematic and fortuitous, which result from a number of different causes and require different treatments in design and maintenance work in order to meet service requirements.

Assume that a given telegraph character is sent continuously over a telegraph circuit and that each repetition of the character is considered perfect as sent. Measurements of the distortion of each of the unit marks (elements) in a large number of successive repetitions of the character at the receiving end will generally indicate that the distortion differs (1) from element to element in a given repetition of the character and (2) from character to character for a particular element in the character. The average of a large number of distortions for a particular element is designated as systematic distortion. The individual departure of the distortion from the average for a given measured distortion is designated as fortuitous distortion. The total distortion of each signal element is the algebraic sum of the systematic and fortuitous distortions and is the amount of deviation between the sent and received signals.

Figure 13 illustrates roughly for 10 repetitions (not enough for a good average in actual cases) the distortion which may affect a marking signal of unit duration as received. The average length of received signal is shown to be 90 per cent of the sent signal, resulting in a −10 per cent systematic distortion which applies to all the repetitions. The departure of the individual distortions from the average varies between repetitions, as shown in the right-hand column, and in the left-hand column the resultant distortion is indicated.

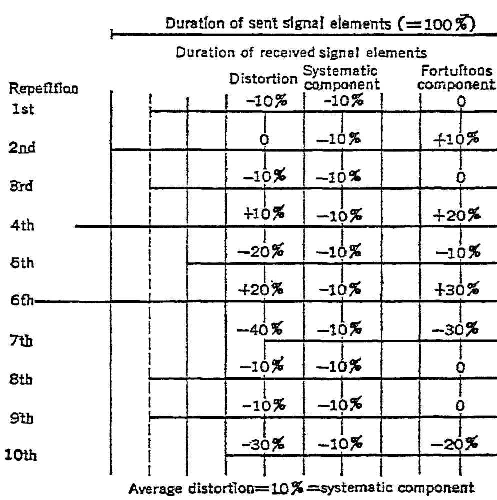


FIG. 13. Systematic and Fortuitous Distortion (Courtesy Bell System)

Systematic distortion may be divided into two component distortions, *bias* and *characteristic*, for the purpose of analyzing and treating the causes of this type of distortion. The nature of these components may be explained by assuming a telegraph system in which marks and spaces are sent by means of currents equal in magnitude but of either positive and negative or negative and positive sign, respectively, as desired.

Bias Distortion. Assume that the systematic distortion, as measured, is due to a higher positive than negative sending-end potential, and that this fact results in lengthening the marks when positive current is used for transmitting marks, and in shortening the marks when negative current is used for transmitting marks. Then, interchanging the functions of the two current conditions employed changes the sign of the systematic distortion but not its magnitude and is called bias distortion, indicating a lack of symmetry in the circuit.

A *marking bias* is called a *positive bias*, and a *spacing bias* is called a *negative bias*; and, since the lengths of the marks and spaces may be indicated in milliseconds (ms), the amount of bias may also be indicated in milliseconds by the formula (M-STD) − (S-MTD) = ms bias, the first two terms being mark-to-space transition delay and space-to-mark transition delay, respectively, and the sign of the result (using this formula) being the sign of the bias. Thus, for a given circuit, if M-STD = 6 ms and S-MTD = 3 ms, the milliseconds bias is +3, indicating that every mark, regardless of length, will be increased 3 ms, and every space, regardless of length, will be decreased 3 ms.

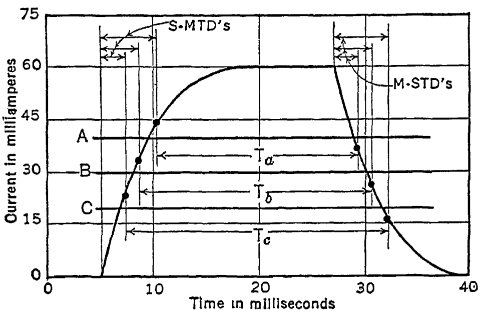


FIG. 14*a*. Effect of Relay Biasing Current on Signal Length (Courtesy Bell System)

Though a given *ms bias condition* is due to such factors as unequal marking and spacing line currents, ground potentials, and improper biased relay adjustments, this condition will remain constant for the transmission of signal reversals (but not random signals) irrespective of the speed of transmission. However, the effect of a given ms bias condition on transmission does vary with the length of the transmitted marks and spaces. For example, in a manual telegraph circuit the dashes (long marks) are about 3 times the length of the dots (short marks), and both dashes and dots decrease proportionately in length with increase in speed of transmission. Assuming first a slow speed, where the dots are 30 ms long and the dashes are 90 ms long, a ms bias of +10 will lengthen the dots and dashes to 40 ms and 100 ms, respectively, the ratio of dash to dot length still being 2.5 to 1. However, if, owing to an increase in speed, the dots and dashes are shortened to 5 ms and 15 ms, respectively, the same +10 ms bias would result in dots 15 ms in length and dashes 25 ms in length. The ratio of dash-to-dot length would then be about 1.7 to 1, and greater difficulty would be experienced in reading the signals than for the usual ratio of 2.5 or 3 to 1.

Figures 14*a* and *b* show the effects on signal length of relay biasing current and line current variations. In the first figure, the line current is held constant at 60 ma, while the normal biasing current (line *B*) is increased (line *A*) and decreased (line *C*). In the second figure, the biasing current is held constant and the normal line current is increased (high line current) and decreased (low line current). It is evident that, in the first condition, increasing the biasing current decreases the signal length and decreasing it increases the signal length; and, in the second condition, increasing the line current increases the signal length and decreasing it decreases the signal length.

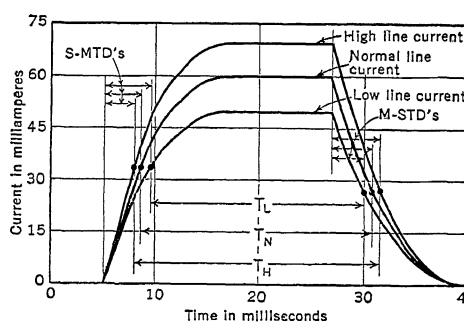


FIG. 14*b*. Effect of Line Current Magnitude on Signal Length (Courtesy Bell System)

Characteristic distortion results from various causes, which are usually different from those associated with bias distortion. Assume a telegraph system, in which the sending battery potentials are equal and opposite in sign and in which the marks and spaces are formed by corresponding currents, equal but opposite in sign. Also, assume that, owing to the characteristics of the given system, the current is slow in building up to the normal mark or space value. If the current does not have time to reach its final value on the short signal elements, the first mark following a long space may be shortened. Under this condition, it is obvious that interchanging the functions of the positive and negative current will not alter either the sign or the magnitude of the resulting distortion, and the distortion is called characteristic distortion, indicating it is a function of the signal combination and fixed characteristics of the system, which causes remnants from a given signal element to affect succeeding elements.

Depending on the speed of operation, telegraph signals are frequently of insufficient duration to permit the line current for a given signal to change from one steady state to the other, i.e., from a maximum positive to a maximum negative line current, or vice versa. In such cases, the transition M-S or S-M will occur while the current is changing, and this is designated as a *changing-current M-S or S-M transition*, both of which are shown in Fig. 15. It will be noted that, for this particular polar telegraph system, the time for the line current to change from one steady-state condition to the other is 33 ms, whereas the duration of the marking or spacing signal is only 22 ms. The net effect on the signal is to shorten it 2 ms, since the transition delay at the start of the signal is 10 ms (subtracts from the signal length) and at the end of the signal is only 8 ms (adds to the signal length). The total current change in either transition is 28 ma (+25 to −3 or −25 to +3); but, if the transmitted signal had been longer than 22 ms, the line current would have exceeded plus or minus 25 ma and at the end of the signal the transition delay

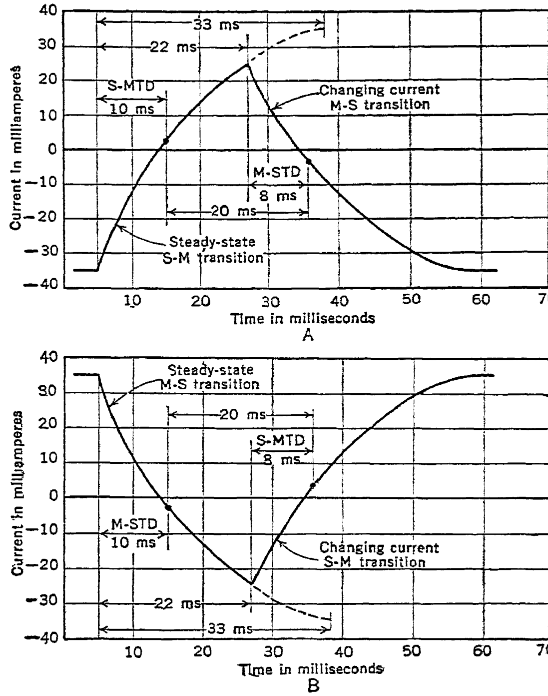


FIG. 15. Changing Current Transitions (Courtesy Bell System)

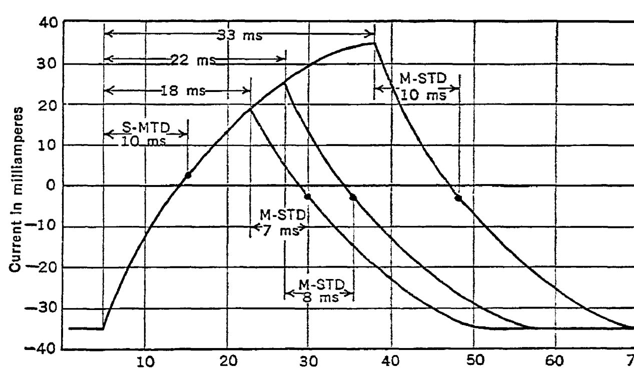


FIG. 16. Characteristic Distortion Effects on Signal Lengths at 40, 60, and 75 Speed Operation (Courtesy Bell System)

would have been greater, up to the limiting delay, resulting from the line current increasing, up to its steady-state value.

Also, if the transmitted signal had been less than 22 ms long, the line current would not have built up to 25 ma by the time the signal ended, and the transition delay at the end of the signal would have been less than for 22 ms. Figure 16 shows characteristic distortion effects on *marking signal length* at 40, 60, and 75 teletypewriter operating speeds. The time required for the line current to change from its maximum negative to positive value, and vice versa, is 33 ms, and the marking signals are 33 ms, 22 ms, and 18 ms long, respectively. Changing current transitions take place for all except the 33 ms signal (40 speed), for which the steady-state current values are just attained. It will be noted that the S-MTD values are the same (10 ms) for all signal lengths, but the M-STD values decrease with decreasing signal length. For similar *spacing signal lengths*, the M-STD values would be the same (10 ms), while the S-MTD values would decrease with decreasing signal length.

The amount of a changing current transition delay is thus dependent on the line current value at the start of the transition, and this line current value depends on the signal length or speed of transmission. The lengths of the received signals are obviously affected by these changing current transitions, and the magnitude of the effect is inversely proportional to the length of the sent signals. Since received short signals are shortened by this effect, the distortion is known as *negative characteristic distortion.* An opposite effect is possible, though not common, and may result, if the line current increases momentarily at the end of each transition to a value exceeding the steady state, owing to circuit characteristics and transient effects. This effect would tend to lengthen the received mark or space signal and would be known as *positive characteristic distortion.*

Some of the principal differences between bias and characteristic distortion are given in Table 1.

Table 1. Differences between Bias and Characteristic Distortion

Type of Distortion Is Affected by	Type of Distortion	
	Bias, ms	Characteristic
1. Length of signal....................................	No	Yes
2. For a given length of signal, whether the signal is marking or spacing.	Yes	No
3. Amount and arrangement of the circuit capacitance, inductance, and resistance.	No, except in neutral operation	Yes
4. Unequal marking and spacing line currents. Change in line current. Change in receiving relay biasing current. Ground potential difference between sending and receiving end.	Yes	No
5. Speed of transmission....................................	No	Yes
6. Usual operating variations, occurring frequently throughout the day, such as voltage fluctuations and relays requiring adjustments.	Yes	Not appreciably

Measurements of systematic distortion, in practice, will usually indicate that both bias and characteristic distortion are present, or the *total measured distortion with the circuit normal is:*

$$S1 \text{ (Total distortion, circuit normal)} = C \text{ (characteristic)} + B \text{ (bias)}$$

and the total measured distortion with the reversed condition is:

$$S_2 \text{ (Total distortion, circuit reversed)} = C \text{ (characteristic)} - B \text{ (bias)}$$

The characteristic component is $(S_1 + S_2)/2$, and the bias component is $(S_1 - S_2)/2$.

Referring to Fig. 17, assume that a repeated signal is being sent over a given circuit, consisting of a marking element 1 unit long and a spacing element 3 units long. If no distortion exists, this signal will be received exactly as sent. However, when measured with the circuit normal, the unit mark is found to be 15 per cent too long, as shown at A. If a measurement is now taken with the line conditions for marking and spacing reversed, this unit mark is found to be 5 per cent too long, as shown at B. By formula, the characteristic distortion C is 10 per cent, and the bias distortion B is 5 per cent, both with signs positive or marking. If, with the line conditions reversed, the unit mark was found to be 5 per cent too short, as shown at B', the characteristic component of the 15 per cent marking distortion (shown at A) would be 5 per cent and the bias component 10 per cent, both positive or marking.

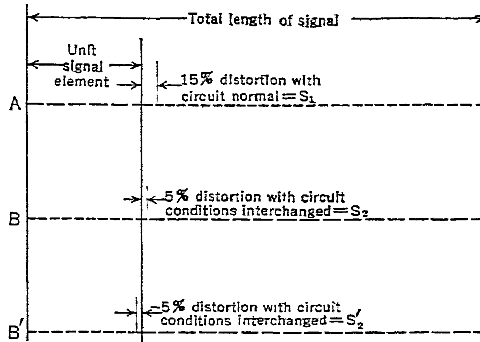
In practice, systematic distortion is usually determined, with the same methods as above, by measuring a repeated signal consisting of a short marking signal followed by a long spacing signal for the normal condition and then a repeated signal consisting of a short spacing signal followed by a long marking signal for the reversed condition.

Though the sign of the final value of characteristic distortion, as computed, is not important, the sign for bias distortion is always important and is indicated.

Fortuitous distortion is caused by such factors as crossfire, power induction, momentary battery fluctuations, line hits, break key operation, and similar effects. This type of distortion acts to alter the received signals by various amounts in an irregular manner. In transmitting miscellaneous signals, the combined effect of all distortion on the displacement of received transitions may result in signals, sometimes called "jitter," because of their rapid variations, or the effect may cause a complete breakdown in signal transmission.

Total distortion, for any given telegraph circuit, is usually a combination of its bias, characteristic, and fortuitous components, and the total distortion determines the quality of telegraph transmission. However, for reasons of design and maintenance, it is usually desirable to determine also the value of these components.

Telegraph distortion is usually given as a percentage of a perfect signal element of unit length. For manual operation, these elements or dots are usually sent at a rate of 12 or 13 dots per second, and their duration (taking into account the dot length interval between each two dots) is about 40 ms. For 60-speed teletypewriter transmission, the signaling rate is equivalent to about 22.7 dots per second, and the duration of the unit signal element, in this case, is 22 ms.



For A and B:

$$\text{Characteristic distortion} = \frac{S_1+S_2}{2} = \frac{15+5}{2} = 10\%$$

$$\text{Bias} = \frac{S_1-S_2}{2} = \frac{15-5}{2} = 5\%$$

For A and B':

$$\text{Characteristic distortion} = \frac{S_1+S_2'}{2} = \frac{15+(-5)}{2} = 5\%$$

$$\text{Bias} = \frac{S_1-S_2'}{2} = \frac{15-(-5)}{2} = 10\%$$

FIG. 17. Characteristic and Bias Distortion (Courtesy Bell System)

TELETYPEWRITER DISTORTION. *Distortion as it affects start-stop teletypewriter signals* must be considered in a somewhat different manner from that for other types of telegraph systems, because teletypewriter system operation differs fundamentally from the operation of other systems. Each teletypewriter operation (the transmission of a character), as previously described in this section, is initiated by the mark-to-space transition at the beginning of the start pulse of the received character. *It is important that each of the succeeding transitions in the character be correctly timed with respect to the first transition. Distortion causes the displacing of succeeding transitions* from their normal positions with respect to the start transition and thus reduces the margin of operation of the teletypewriter.

Bias distortion usually affects both the beginnings and ends of the received signal elements. However, the teletypewriter receiving mechanism starts on a *mark-to-space transition* (for each character received), and this transition is also affected by the same bias, so that succeeding mark-to-space transitions in the character will not be displaced with respect to the start transition, owing to bias. The *space-to-mark transitions* will be displaced with respect to the start transition by an amount equal to the total bias. With *marking bias,* all space-to-mark transitions will be uniformly displaced toward the start transition, whereas with *spacing bias* they will be uniformly displaced away from the start transition.

Characteristic distortion may displace both the received space-to-mark and mark-to-space transitions with respect to the start transition, depending on the signal combination, and recurs for the same signal combination. It may affect both ends of the teletypewriter orientation range, and when miscellaneous characters are received a distinction cannot be made between the characteristic and fortuitous components of the distortion.

Fortuitious distortion displaces miscellaneous received transitions by various amounts in a random manner, regardless of the signal combination. These effects may result in errors in received characters or, if severe, in complete circuit failure.

The total distortion is the displacement of a received transition from its correct time of occurrence, and it is equal to the *algebraic sum of the fortuitous distortion and the systematic distortion.* The *total distortion* determines the margin of operation of the receiving teletypewriter, and it is a *measure of the transmission quality of the received signals*

In teletypewriter operation, not only signal distortion takes place during transmission, but also *mechanical variations occur in the teletypewriter mechanisms* which affect the quality of signal transmission. The mechanical operation of the teletypewriter is described briefly in article 7 of this section, but, for the purpose of discussing distortion as it affects teletypewriter operation, some of the mechanical features of the teletypewriter must be referred to in this article.

Figure 18 shows some of the principal mechanical units of the selecting arrangement of a teletypewriter which determines the instant, in the duration of a given received signal element, when the actual selection takes place. For example, if the signal element being received is marking, the selection must be made at some instant during the time this

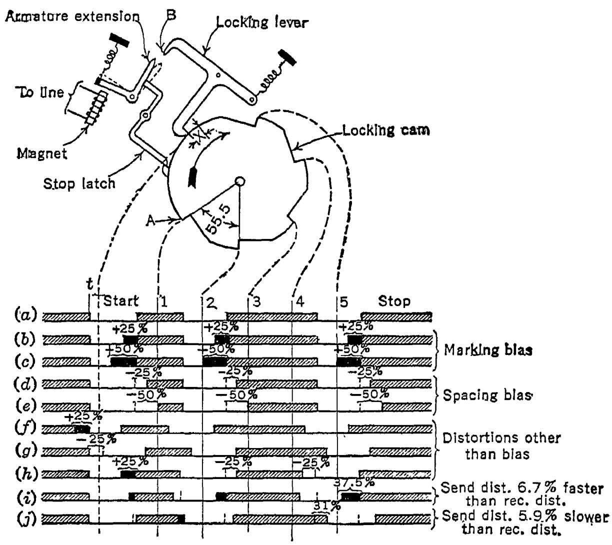


FIG. 18. Distortion Effects on the Selection of Teletypewriter Signals (Courtesy Bell System)

particular element is being received, in order that the code bar, corresponding to this selection, will be properly positioned and the character of which this element is a part will be correctly recorded.

The selecting mechanism includes: (1) the line magnet, which is operated by marks and non-operated by spaces; (2) the armature and armature extension, which are held in the position shown during marking and are released to the dotted position during spacing; (3) the stop latch, which stops the locking cam rotation after each revolution; (4) the locking lever, whose point *B*, during rotation of the locking cam, engages one side or the other of the end of the armature extension, depending on the position of the extension, thereby locking the extension momentarily, which results in the code bars being properly positioned; and (5) the locking cam, which is driven by a friction clutch and which, in a single revolution, causes the locking lever to be positioned five times, once for each of the five signal elements received, during the transmission of one character.

In the idle condition, the received signal is always marking and the magnet is energized, but the selector cam assembly is held stationary by the stop latch until the start pulse (always spacing) is received, when the stop latch is released and the cam starts rotating. The speed of rotation and the starting position of the selector cam assembly are normally so adjusted that the first depression on the locking cam (point *A*) will arrive at the locking lever point (which rides the cam) at about the instant the middle of the first selecting signal element is being received. At this instant the locking lever point (riding the cam) moves into the cam depression, and the lever point *B* moves forward to lock the armature and its extension in one of the two positions they may occupy at that instant. The position of the armature extension determines which of the two line conditions, marking or spacing, will be recorded for the signal element being received, and the position of the corresponding code bar. As the cam rotates, a similar selection takes place as the locking lever enters each of the five depressions on the cam and the corresponding code bars are positioned in succession. The final arrangement of the five code bars results in the selection of one and one only type bar, which, when actuated immediately after completion of one revolution of the locking cam, prints the sent character on paper. The final pulse of the train of signal elements (for each character received) is always marking and somewhat longer (1.42 units) than the preceding six elements (1.0 unit each). During this pulse, the stop arm on the receiving selector cam assembly strikes the stop latch, and the assembly is then held stationary until the next start pulse is received for the next character.

When a start pulse is received, a small increment of time will elapse before the selector cam assembly attains full speed, owing to such factors as the inertia of moving parts and clutch slippage. This delay is compensated for by slightly decreasing the distance between the point at which the locking lever rests on the locking cam, and point A on the cam, from what the distance would be if these factors were not present. This adjustment is represented by the distance x, as shown on the cam surface in Fig. 18.

The teletypewriter is usually equipped with an *orientation or range finder device*, which permits rotating the stop latch with respect to the locking lever, thus changing the time of selection with respect to the start signal. The range finder moves the stop position either forward or backward and has a pointer which moves along a scale, the scale being calibrated in percentage (0 to 120) of a unit signal element. *If the pointer is moved toward the lower-numbered part of the scale, the time between the start and selecting points is reduced* and the time of selection is advanced toward the beginning of each selecting element. *If the pointer is moved toward the higher-numbered part of the scale, the time between the start and selection is lengthened* and the time of selection is moved toward the end of each selecting element. For an ideal teletypewriter, whose mechanism acted instantly and selected exactly the corresponding instant of each signal element, the range finder could be moved over a range of 100 per cent, if perfect signals were received, without causing errors in the received signals, as shown by (*a*) in Fig. 18. Moving the range finder, in effect, shifts the solid vertical lines with respect to the signal elements, and for (*a*) the time of selection could be changed, without error, by ± 50 per cent. Practically, teletypewriter specifications require the overall range to be at least 72 per cent without error for perfect signals.

Distortion in teletypewriter signals, as previously stated, is usually some combination of bias, characteristic, and fortuitous distortion. Theoretically, if bias exceeds 50 per cent in (*c*) and (*e*) of Fig. 18, errors will result in the recorded signal, but, from a practical standpoint, the bias tolerance is of the order of ± 40 per cent with perfect received signals because of allowances that must be made for other variations.

Internal bias may exist in a teletypewriter as the result of such factors as improper adjustment of the line relay or receiving magnet. This bias reduces the orientation margin more when receiving perfect signals than when receiving signals with a bias equal in magnitude but opposite in sign. In order to minimize this bias effect, the range finder should be set at the point where signals having equal marking and spacing bias just cause errors.

The effect of shortening the start pulse by 25 per cent, (*f*) in Fig. 18, is equivalent to lengthening the stop pulse and retarding the points of selection for the succeeding signal elements by the same amount. The effect of shortening the end of the stop pulse by 25 per cent (*g*) is equivalent to advancing the points of selection for the succeeding signal elements by the same amount. The effect of advancing the beginning of element 1 and of retarding the beginning of element 3 and the ending of element 4, as shown by (*h*), will not result in errors in the received character if the range finder is set at its midpoint. The effect of speed variations where the sending machine is faster than the receiving machine is illustrated by (*i*), and where the reverse is true is shown by (*j*). The most probable error, in both cases, would be in the proper selection of element 5, since, in the first condition, part of the stop pulse is received on the 5 position, and, in the second condition, either element 4 is extended into position 5 or element 5 is so delayed in starting that it would not be properly received in its normal position. The effect on teletypewriter margin of speed variation is mostly at one end or the other of the orientation range, depending on the relative speeds of the sending and receiving machines. Though the speed differences in Fig. 18 are shown large for purposes of illustration, actually these differences are normally about $1/2$ per cent or less.

The orientation range limits are determined by the various distortions present for a given teletypewriter circuit. It is not possible to judge these distortions *quantitatively* by the limits obtained when more than one type of distortion exists. However, assuming low machine bias, these limits do give a good indication of the *quality* of the received signals. If the limits are reasonably definite, some fixed distortion, such as bias or speed difference, is generally present, while, if there is a certain range at each limit over which certain characters are consistently in error, characteristic distortion is indicated. If there is a range over which errors occur, but not consistently on certain characters, fortuitous distortion is most likely present.

TELEGRAPH SYSTEMS

By John D. Taylor

Telegraph systems employ various types of telegraph equipment and connecting mediums and various methods of transmission. Direct-current telegraph systems are used extensively in land wire operation and in conjunction with radio channels, while low- and high-frequency carrier telegraph systems utilize both land wire lines and radio channels as mediums of transmission. Radio telegraph systems make use of radio channels between radio transmitting and receiving equipments and usually wire line extensions between such equipments and the telegraph circuit terminals.

6. DIRECT-CURRENT SYSTEMS

Open-wire telegraph channels are generally obtained by using bare telegraph wires. They may also be provided by simplexing or compositing open-wire telephone circuits, as shown in Figs. 1 and 2, respectively.

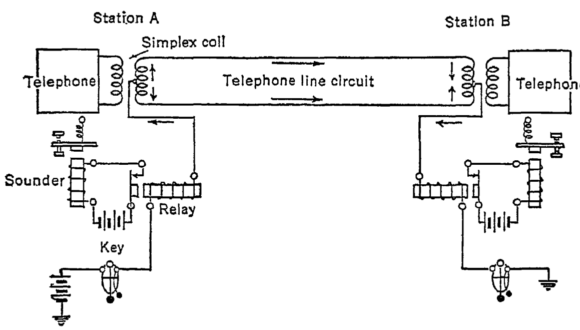


Fig. 1. Telegraph Circuit on Simplexed Telephone Circuit (Courtesy Bell System)

The simplex arrangement is shown in Fig. 1. The arrows represent the telegraph line currents, which divide equally at the junction of the two halves of the line windings of the repeating coil at A and travel over both line wires of the telephone circuit to the repeating coil at B, where they again combine at the junction of the two halves of this coil before passing to Station B telegraph equipment. This equal division of the current is possible only if the two halves of the line side of each coil are identical electrically and the two line wires have identical electrical characteristics. To the extent that these conditions are not met, the current will not divide or combine equally at the coils and a residual induced current will flow in the drop windings of the repeating coils, causing interference (*Morse thump*) in the telephone circuit. In practice, Morse thump becomes objectionable only when faults occur.

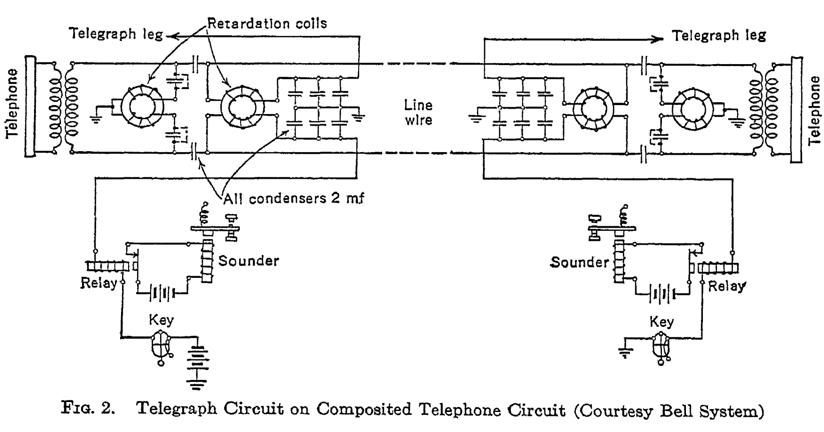


Fig. 2. Telegraph Circuit on Composited Telephone Circuit (Courtesy Bell System)

The composite arrangement is shown in Fig. 2. One grounded telegraph channel is obtained from each line wire of the telephone circuit by connecting composite sets at each terminal of the circuit in such a manner as to maintain the balance of the circuit and avoid excessive telegraph current flow in the line and drop windings of the repeating coils, which current would cause Morse thump. The equipment on the two sides of the circuit must be well balanced and the line wires must be closely identical electrically to prevent telegraph signals from interfering with the telephone service. With this arrangement each telegraph channel is independent of the other and is usually so operated.

The composite set has a retard coil in series with the telegraph leg and capacitance between the leg and ground to prevent sudden changes in signal current value being impressed on the line wires and causing current surges (clicks) in the telephone circuit. Also, the series condensers in the telephone circuit prevent direct current from reaching the repeating coils and assist in reducing clicks, while the retard coil-condenser bridge on the drop side of the series condensers functions to prevent *crossfire,* a condition where the telegraph signals on one wire of the circuit induce potentials on the other wire that interfere with the telegraph signals over it.

A schematic diagram of a neutral telegraph circuit is shown in one form in Fig. 6*a*, article 4. This circuit operates with current flowing in either direction for the marking condition and no current flowing for the spacing condition. Both sending and receiving relays usually operate local sounder circuits, which produce the recognized audible dots and dashes of the Morse or Continental telegraph codes by the armature of the sounder striking its front and back contacts, corresponding to the marking and spacing conditions. Neutral telegraph operation is also employed in teletypewriter and teleprinter service on many of the shorter circuits.

The receiving station may break the circuit (stop the transmission of signals) by opening the sending key, which silences both sending and receiving sounders and indicates to the sending station that the receiving station desires to send signals.

A number of intermediate stations may be connected in series in the single wire line, the number depending on the sensitivity of the line relays, the battery voltages applied, and line conditions, such as resistance and leakage.

The single line repeater, a schematic diagram of which is shown in Fig. 3, may be employed over long single wire circuits. This repeater functions to receive weak signals from the line from either direction and repeat them with normal voltage to the line in the other direction. The signal transmission circuit through the repeater is shown by heavy lines. Operation of the *west station key causes relay A to repeat signals to the east circuit,* and operation of the *east station key causes relay A′ to repeat signals to the west circuit.*

Opening either station key interrupts current flow in both east and west lines, and, unless auxiliary circuits were provided in the repeater,

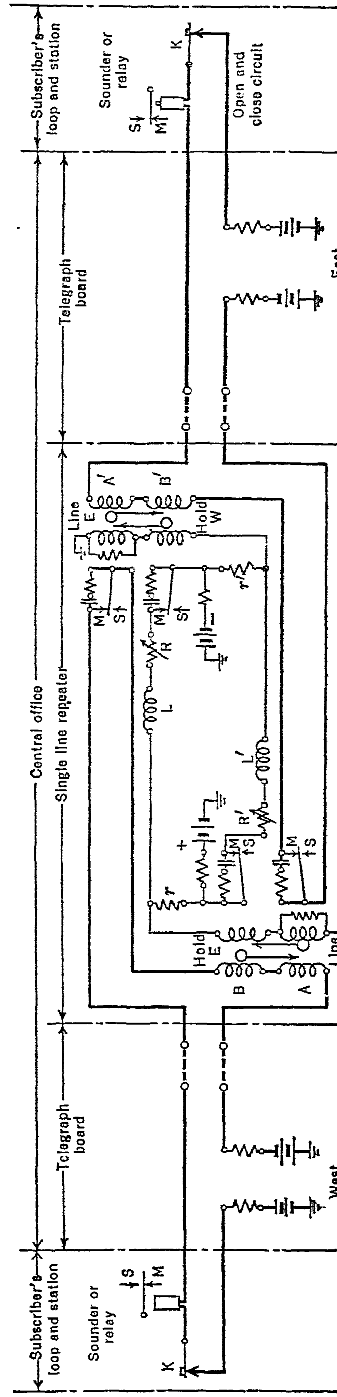


FIG. 3. Single Line Repeater Using Polar Relays—Connected to Open and Close Type Circuits (Courtesy Bell System)

both circuits would open and remain open, and the system would be inoperative. To prevent this condition, a biasing and a locking circuit are provided (see Fig. 3) for each direction of transmission. Each biasing circuit operates the armatures of the line and control relays, with which it is associated, to spacing, when the circuit through the line windings of these relays is opened during the transmission of signals. However, the locking circuit is so designed that, for *east-to-west transmission* with the east circuit open, the biasing current reverses direction in the biasing windings of the line west relays and the armatures of these relays are held on marking, thus maintaining the east circuit closed at the repeater, regardless of the position of the east station key. The locking circuit functions oppositely for *west-to-east transmission.*

A break feature (see Fig. 3) is provided, so that the receiving station may interrupt the sending station, as desired, by opening the receiving-station key.

The one-way polar telegraph circuit, shown in Fig. 9, article 4, usually employs −130 volts for marking and +130 volts for spacing signals to a distant polar receiving relay. This circuit operates from the central office to an outlying point and is used in certain cases where a one-way service only is required, such as in the transmission of news copy.

Two-path polar operation consists essentially of two one-way polar circuits operating in opposite directions.

Polarential operation provides for true polar operation from the central office to an outlying point and a modified polar operation in the opposite direction. The advantages of polar over neutral operation (particularly self-compensation of line leakage) are thus obtained with relatively simple equipment arrangements at the outlying point.

Figure 4 shows a type A polarential telegraph circuit, in which true polar signals are transmitted from the central office, and ground and −130 volt battery (in series with a

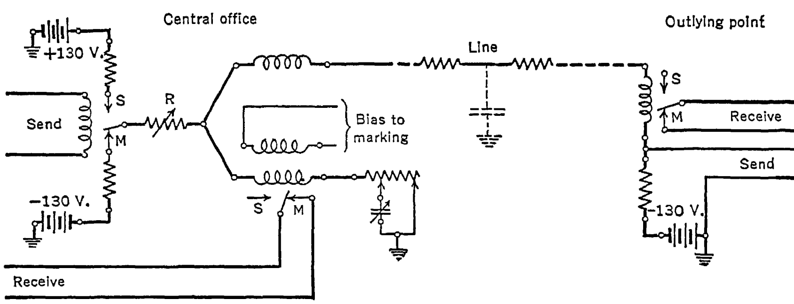


FIG. 4. Type A Polarential Telegraph Circuit (Courtesy Bell System)

total added resistance of 990 ohms) are applied in transmitting marks and spaces, respectively, from the outlying point. In this case a repeater is not used at the outlying point, as is more commonly done in type A operation.

During transmission from the central office, the circuit is closed to a direct ground at the outlying point through the sending contacts (in this case, the contacts of a teletypewriter) while the central-office receiving relay is held on its marking contact by a marking biasing current. This current is adjusted to one-half the effective spacing current when a spacing signal is being sent from the outlying point. During transmission from the outlying point, the ground, applied for a marking signal, has no effect on the central-office receiving relay, provided that the line and artificial line are balanced at the duplex set, and this relay is held on its marking contact, while negative battery, applied through the resistance for a spacing signal, produces an effective spacing current in the central-office relay. This current is the net result of current flowing in the central-office relay windings from (1) the outlying-point battery and (2) the central-office battery, due to the duplex balance being upset by the 990-ohm resistance in series with the outlying-point battery.

The variable resistance R at the central office is adjusted so that the spacing line battery at the outlying point is higher than the potential applied to the apex of the duplex at the central office, thus insuring line current reversal when a spacing signal is transmitted from the outlying point and home copy if a teletypewriter is employed.

Figure 5 shows a simplified type B polarential telegraph circuit, in which true polar signals are transmitted from the central office, and ground and +130 volt battery are used in transmitting marks and spaces, respectively, from the outlying point. This circuit is more nearly self-compensating for line leakage than the type A circuit.

When transmitting from the outlying point over a dry (no leakage) line, the marking line current does not affect the central-office receiving relay, if the balancing network at the central office exactly balances the line electrically.

For a spacing signal from the outlying point, aiding positive battery, applied to the line, results in an effective spacing current $E/(R_1 + R_2)$, where E is the outlying battery potential and $R_1 + R_2 = R_L$ are as shown in Fig. 5. The marking biasing current in the central-office receiving relay is adjusted to a value equal to $E/2R_L$.

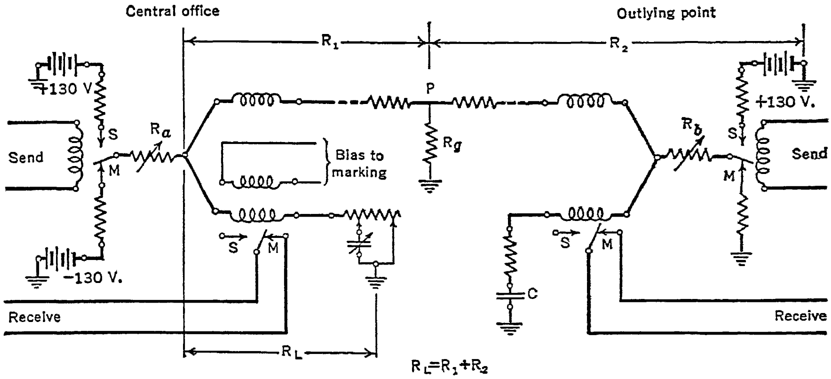


FIG. 5. Type B Polarential Telegraph Circuit (Courtesy Bell System)

During transmission over a *wet line* with leakage R_g (shown at P on the line), R_a (at the central office) may be adjusted to maintain the potential at the apex of the relay windings at the same value as for the dry condition when transmitting a spacing signal from the outlying point. Thus, as a result of the compensating effects of the R_a adjustment, the received signals at the central office are not affected. If R_1 is greater than $2R_2$, complete leakage compensation is not possible, unless the central-office battery voltages are made higher than the outlying battery voltage or other compensation is provided. For this purpose, R_b is provided for maintaining R_1 less than $2R_2$.

The bridge arm, with a resistance in series with condenser C, is provided at the outlying point to neutralize reverse current surges through the receiving relay winding when the outlying sending relay armature moves from space to mark. Such surges would tend to cause false breaks (kick off) of the receiving relay and mutilation of the home copy.

Metallic telegraph circuit operation generally utilizes telegraph cable pairs or open wires, or, when telephone facilities are involved, it is customary to composite them to secure the necessary telegraph channels. The avoidance of interference between telegraph

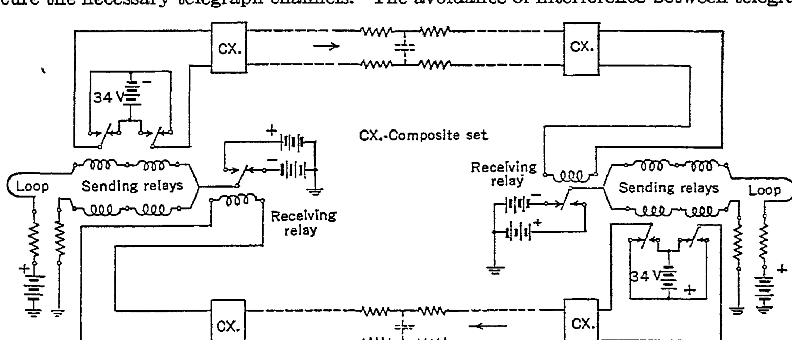


FIG. 6. Four-wire Metallic Telegraph Circuit (Courtesy Bell System)

and telephone circuits and of crossfire between telegraph circuits in cables generally requires that the telegraph current values be comparable to those on the telephone circuits and that metallic circuit operation only be employed. Such operation may employ two wires or four wires, pairs, side circuits, or phantoms.

In four-wire operation, as shown in Fig. 6, separate paths are employed for the two directions of transmission, and artificial balancing lines are not required, as in two-wire operation. These differences generally improve transmission over that of the two-wire arrangement, since balance between the conductors and an artificial line is not a factor. Owing to the use of four composited channels, as compared to two such channels for two-wire operation, the possibilities for conductor and equipment troubles are greater with four-wire operation. In general, four-wire telegraph circuits may be superposed on two-wire telephone circuits of lengths ranging from about 500 to 1000 miles, or on longer four-wire telephone circuits. Very long four-wire telegraph circuits are not composited throughout their entire length because of prohibitive low-frequency delay distortion introduced by the composite sets.

Duplex systems may be of the earlier *bridge* type or the later, more commonly used differential type. Both these systems employ arrangements of telegraph apparatus for terminating telegraph circuits at the central office which permit the simultaneous transmission of telegraph signals in both directions over a single wire with ground or metallic return and without the signals in one direction interfering with those in the opposite direction.

Since bridge polar duplex operation is rapidly being discontinued and is expected to be of little interest in the future, its description is omitted from this handbook. However, occasional reference is made to it for purposes of comparison in the following paragraphs, which relate to differential polar duplex operation.

Figure 7 shows schematically the *differential type* set, arranged for full duplex operation. The receiving relay has two equal windings, one being connected in the real line and the

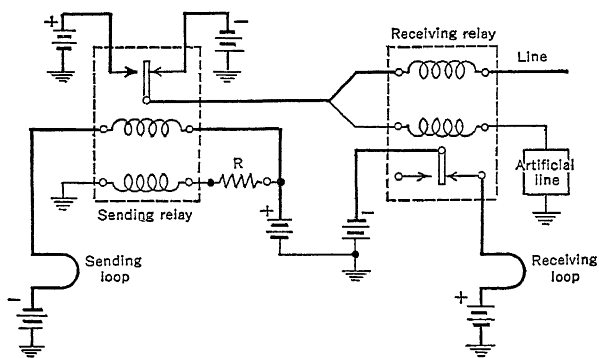


FIG. 7. Terminal Differential Duplex Set Arranged for Full Duplex Service (Courtesy Bell System)

other *differentially* in the artificial line circuit. Sending current from the station battery divides at the apex of the receiving relay windings and flows equally in the real and in the artificial line circuits, assuming these are balanced. Since the receiving relay windings are differentially connected, the station receiving relay is not affected by this current flow. However, current from a distant station does operate the receiving relay at the home station, since it flows through the windings of this relay in a series aiding relation from the line through the balanced arms and the artificial line to ground. Similarly, the receiving relay operates at the distant station from incoming line current from the home station. Under these conditions full duplex operation is attained.

In the differential-type duplex set, the sending relay has both an operating and a biasing winding. With the sending loop key closed, the magnetic fields produced by these two windings are opposing, but a resistance R in series with the biasing winding limits its current to about half that in the operating winding and the armature is accordingly held to the marking contact (negative). When the loop key is opened, only the biasing winding is effective and the armature moves to its spacing contact (positive).

Duplex systems, employing the ground for one side of the circuit, are affected by variations in line constants due to temperature and humidity changes, but the balancing artificial lines associated with each duplex set usually may be adjusted to compensate for such variations. Duplex systems provide all the advantages of polar operation in both directions of transmission, but they also necessitate higher-grade supervision and greater maintenance than some of the simpler systems, such as in neutral operation.

The differential duplex system has largely superseded the older bridge duplex system because of several advantages of the differential over the bridge type. The differential system includes a simple differential duplex repeater for use at intermediate points on long telegraph circuits; in the bridge system, two terminal sets, requiring considerably more equipment, must be used at such points. Intermediate differential repeaters may be employed on circuits having bridge polar type sets at their terminals, or bridge polar intermediate repeaters may be used on circuits having terminal differential sets at their terminals.

One of the most important advantages of the differential over the bridge system is the ability of its polar relays to respond rapidly to signals, particularly at the higher operating speeds. These relays, in addition to being more sensitive than the relays ordinarily used in bridge polar sets, have a special third winding which forms part of a *vibrating circuit* and materially increases the relay response. The vibrating circuit, shown in Fig. 8, includes the third winding of the receiving relay, whose armature is connected to the midpoint of the winding. One outer terminal of this winding is grounded through a condenser, and the other outer terminal is grounded through a resistance. Batteries of opposite polarity are connected between ground and the relay contacts.

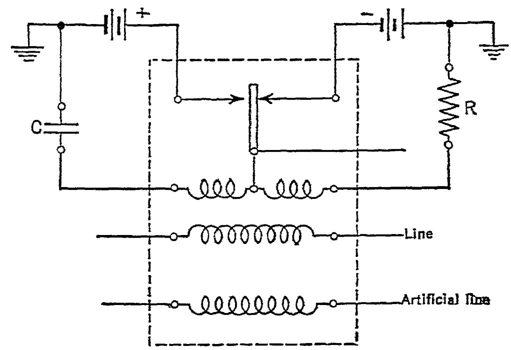


FIG. 8. Vibrating Circuit of Differential Relay (Courtesy Bell System)

Assuming that no current is flowing in either the line or artificial line windings of the receiving relay, the relay armature moves back and forth continually between the two contacts. However, with current flowing in the line and artificial line windings in normal operation, the armature does not vibrate freely as before, but the vibrating circuit aids in speeding up the armature action, as the received line signals cause it to move between its marking and spacing contacts. The free speed of vibration is governed by the parameters C and R, which are usually adjusted so that the frequency of vibrations is slightly greater than the line signal frequency.

Half duplex operation, which provides for operation in only one direction at a time, requires a means for breaking the circuit from either terminal and for utilizing one loop for both sending and receiving signals. Either the bridge- or differential-type system may be employed for this service, but each requires a different circuit arrangement of the duplex set from that required for full duplex operation. In the bridge-type system, the principal additions required are a control relay, holding coil for the pole changing relay, and a repeating sounder; for the differential-type system a control, break, and neutral relay are added. The change from full duplex to half duplex, or vice versa, for either system is readily accomplished by means of switches provided with the duplex sets.

Figure 9 shows a schematic diagram of a terminal differential duplex arranged for half duplex service. The control relay functions on incoming signals to prevent the sending

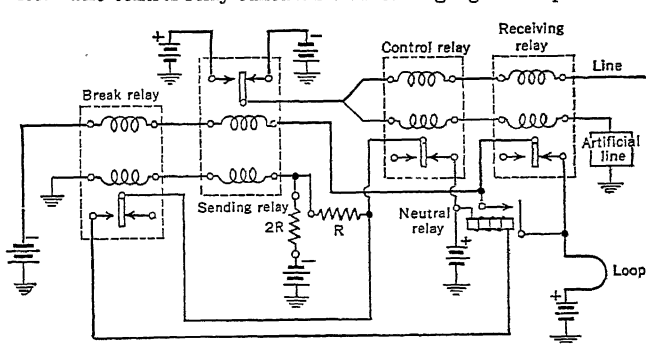


FIG. 9. Terminal Differential Duplex Set Arranged for Half Duplex Service (Courtesy Bell System)

relay from operating when the loop is opened and closed by the receiving relay. Two oppositely poled batteries are connected to the biasing winding of the sending and break relays, and the resistances R and $2R$ are so adjusted that, with the control relay contact closed, the positive (spacing) battery will be in control and the sending relay will operate in the same manner as in full duplex operation.

When signals are being received from the line, a marking signal closes the loop, and current flows from the loop through the sending relay operating winding. When a spacing signal opens the local loop at the receiving relay contacts, no current can flow through the sending relay operating winding. However, the control relay also operates in unison with the receiving relay (since their windings are in series), opening the positive (spacing) battery at its armature contact and permitting negative (marking) battery to take control of the sending and break relays through resistance $2R$. Thus, the sending relay armature is held on its marking contact for either incoming marking or spacing signals.

If the local operator wishes to *break* the circuit while incoming signals are being received, the local loop circuit is first opened by opening the sending key. The next spacing signal received with the loop open would be ineffective, since the signal would normally open the loop at the receiving relay contact. The next marking signal received with the loop open would operate the receiving and control relays, and positive (spacing) battery would be applied to the biasing windings of the sending and break relays through the control relay armature contact and resistance R. Since there is no current flowing through the operating windings of the sending and break relays with the loop key open, both relays operate to spacing. Positive battery is then applied to the line through the sending relay spacing contact, which results in a break signal to the distant operator.

The break relay functions to insure a continuous break signal to the line as long as the local loop key is open, regardless of subsequent signals received from the line. Assuming that one of these signals is a spacing signal, the receiving and control relay contacts will open, which would permit negative (marking) battery to take control of the sending and break relays and marking battery to be applied to the line through the sending relay marking contact if a secondary circuit was not provided. Marking battery applied to the line would interrupt the break signal. However, positive current is maintained through the biasing windings of the sending and break relays as soon as the break relay operates to spacing (which it would do on the preceding incoming marking signal), since positive battery is then applied through the neutral relay, spacing contact of the break relay, resistance R, and the biasing windings of the sending and break relays to ground.

The neutral relay in this secondary circuit functions to prevent the possibility of the circuit becoming inoperative owing to the sending and break relays at both terminals becoming simultaneously operated to spacing. Under this condition, neither operator could regain control of the circuit, since the loop circuits at both terminals would be open at the receiving relay contacts. Operation of the neutral relay which occurs when the control relay is opened after the break relay is operated to spacing short-circuits the receiving relay contacts, so that the sending and break relays will be operated to their marking contacts when the associated loop key is closed.

The upset duplex method of operation between a central office and an outlying point employs polar transmission generally from the central office to the outlying point and

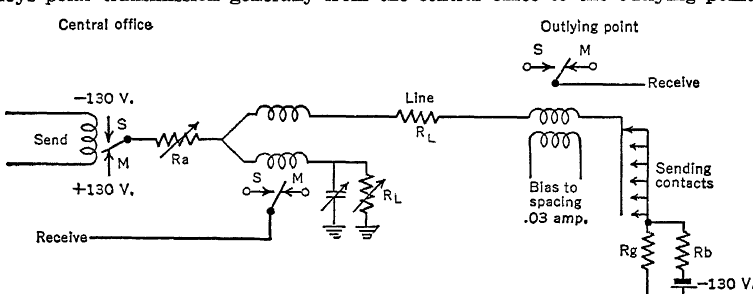


FIG. 10. The Upset Duplex Method in Telegraph Operation (Courtesy Bell System)

neutral transmission in the opposite direction. Transmission in either direction is affected by line leakage, and transmission to the central office is subject, in general, to the same limitations as for a neutral circuit employing a single line repeater. Because of polar operation in one direction and the application of wave-shaping units at the outlying point only, this method offers an improvement in operation over the neutral circuit with a single line repeater.

Figure 10 shows the upset method of operation, employing a duplex set, arranged for half duplex operation at the central office and a teletypewriter (neutral operation) at the outlying point. The teletypewriter normally operates on 60 ma of line current in an open and close loop circuit, and its line relay requires a spacing biasing current of 30 ma. The algebraic sum of the marking and spacing line currents while sending to the outlying point must equal 60 ma, in order to meet the requirements for operation under the upset method. Also, when receiving at the central office (neglecting the effects of series inductance and bridged capacitance on the line), the effective spacing current in the receiving relay with the outlying point key open must equal the effective marking current in this relay with the outlying point key closed. In practice, it is generally necessary to adjust these spacing and marking currents in the receiving relay, since the inductance and capacitance effects cause unsymmetrical wave shapes and cannot be neglected.

The **duplex repeater** is employed at intermediate points in long circuits to maintain proper operating currents in duplex telegraph systems. The distance between repeater points depends on a number of fixed and variable factors, such as the type of line facilities, interference from extraneous sources, line leakage, coordination of transmission levels, types and quantities of central-office equipment involved, line operating speeds, and maintenance considerations. Figure 11 shows schematically an intermediate differential duplex repeater circuit for grounded operation.

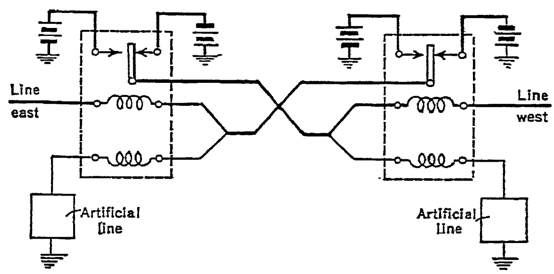


Fig. 11. Schematic of Intermediate Differential Duplex Repeater (Courtesy Bell System)

For 60-speed start-stop teletypewriter operation, the maximum lengths of single composited line sections of 104 mil and 165 mil copper line wire, over which it should be practicable to operate either the differential duplex or two-path polar systems without intermediate repeaters, will be roughly within the range of 170 to 300 miles and 250 to 450 miles, respectively. For any given section, the maximum length depends on a number of variables, such as the types and quantity of terminal and intermediate equipment used, the line conditions, and maintenance schedules and type of personnel.

Where thump, flutter, and crossfire considerations permit, 60-speed differential duplex or two-path polar operation is usually feasible in cable over a composited 13 or 16 gage wire with crossfire neutralizing networks or over a 19 gage simplexed pair or phantom for distances up to about 100 miles.

The quadruplex system permits four messages, two in each direction, to be transmitted simultaneously over a single grounded circuit. One transmission in each direction is secured in the same manner as has been previously described under the heading Duplex Systems, article 6. These two channels and the associated sending and receiving apparatus are called the *polar side of the quad.* The two additional transmissions (one in each direction) are obtained by varying the strength irrespective of the direction of the line current, and by receiving the signals, thus sent, on a neutral relay, which responds to impulses of large amplitude irrespective of polarity but remains unaffected by impulses of smaller amplitude. These two channels and the associated equipment are designated the *neutral or common side of the quad.*

The schematic circuit of the differential quadruplex is shown in Fig. 12.

Normally a resistance is in series with the pole-changing relay, and a leak to ground is connected to the line, to preserve the proper ratio between the marking and spacing current of the common side. When this resistance is short-circuited, and the leak resistance removed by the contacts of the transmitting relay, in response to the operation of a sending key which controls it, the line current is increased to its maximum value. The neutral relays in the receiving circuits are adjusted so that they will respond to the stronger signaling currents but will be unaffected by the weaker ones. Momentary release of the neutral relay at the instant the direction of line current is reversed by operation of the distant pole-changing relay is prevented by the discharge of a condenser through a special holding winding.

The normal operating currents in a quadruplex set are 10 to 20 ma for the polar relays and 30 to 60 ma for the neutral relays. A ratio of 1 : 3 or 1 : 4 between the spacing and marking current is required to insure good operation.

Although employed quite extensively at one time, the quadruplex system is now used but little, because it was found impossible to maintain uninterrupted operation of the

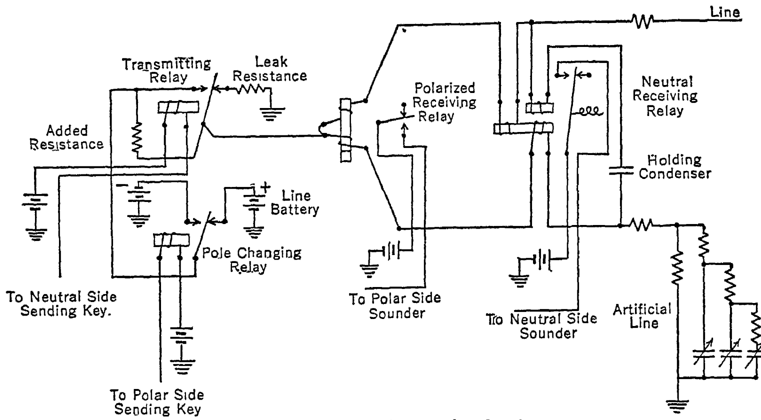


FIG. 12. Differential Quadruplex Set

common side during periods when damp weather or other causes lowered the line insulation sufficiently to reduce the ratio between the operating and non-operating currents.

The common side of a quad is also more susceptible to interruption by inductive interference from adjacent circuits.

The bridge principle, described under duplex systems, may also be utilized in a quadruplex system.

7. AUTOMATIC TELEGRAPH SYSTEMS

Start-stop systems, employing the *teletypewriter* and *teleprinter* equipment (similar type units of the Bell System and Western Union Co., respectively) for sending and receiving telegraph signals, are so named because of the method of operation. This equipment usually employs the seven-element code, of which the first element is a start pulse, the seventh element is a stop pulse, and the other five elements represent the character. Thus, the sending and receiving units are synchronized after each group of seven pulses (elements) is received, constituting the transmission of a character.

Usually, the teletypewriter or teleprinter installation at a subscriber location consists of an electromechanical unit, with or without a keyboard resembling that of a standard typewriter, and having a typing mechanism for printing received messages on a paper page or tape. At receiving-only stations, the sending keyboard is not required. In some cases, the subscriber may perforate a tape, using a *keyboard-equipped perforator*, the perforations representing coded characters of the desired message. This tape may at the same time or later be passed through a *tape transmitter*, which, being connected to the line, sends pulses over the line corresponding to the characters of the message. Also, incoming code pulses may be received by a *reperforating machine*, which perforates a tape with the five-element code representing the received characters. Figure 13 shows a sample of perforated tape.

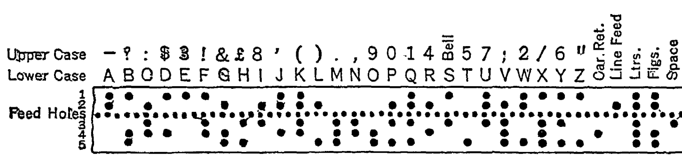


FIG. 13. Teletypewriter Code Perforated in Tape

The teletypewriter sending unit now in general use consists principally of a keyboard with key levers extending over five notched selector bars, and a start-stop mechanism of driving and driven shaft, universal bar, cams, eccentrics, levers, and pawls. Figure 14 shows the general mechanical arrangement of the start-stop mechanism.

Figure 15 shows the details of the key lever, which, when pressed down, positions the selector bar, which, in turn, moves the *locking latch head* forward or back. This latch will either prevent the *contact lever* from closing the *transmitter contacts* (latch head forward) or permit the contact lever to close these contacts (latch head back), *when the associated* *selector cam depression arrives at the proper projection on the contact lever.* Thus, as a key lever is depressed to send a particular character, the *universal bar* is moved down, causing the mechanism (Fig. 14) to function and the selector cams to start rotating. At the same time, the five selector bars and locking latches are positioned by the key lever, and each

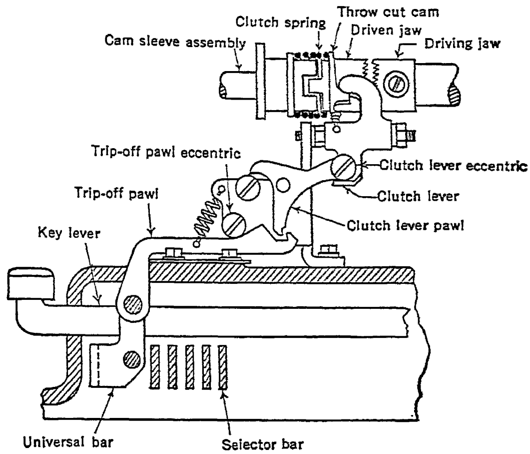


FIG. 14. Start-stop Mechanism of Teletypewriter Sending Unit (Courtesy Bell System)

of the five contact levers is either locked or left unlocked, so that, as the selector cams make one revolution, each set of transmitter contacts does not close or does close. If the contacts do not close, a spacing signal (no current) is sent over the line, and if the contacts close a marking signal (current) is sent over the line.

Release of the key lever after it has been fully depressed causes the driven jaw to be thrown out of engagement with the driving jaw upon completion of one revolution of the

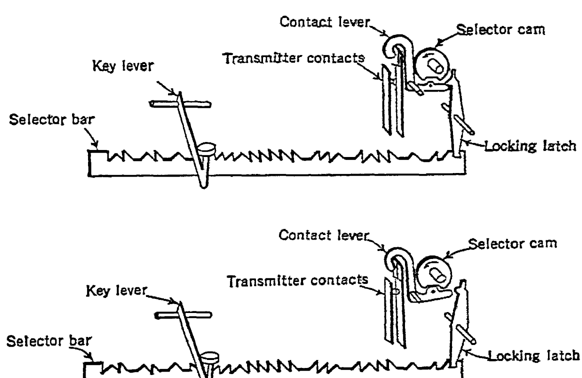


FIG. 15. Positioning of Transmitter Contacts by Operating Key Lever of Teletypewriter Sending Unit (Courtesy Bell System)

cams, and the rotating movement stops. The mechanism is now ready to send the next character.

The teletypewriter receiving mechanism now generally employs a single selector magnet and a group of six rotating cams, so spaced angularly on a shaft that each cam functions at the instant the corresponding signal pulse is being received.

Figure 16 shows the mechanical arrangement for translating the selector magnet operations into the positioning of the code bars. When the open pulse is received, the magnet armature releases. This operates a latch (not shown), allowing the cam shaft to rotate. The spacing of the cams on the shaft is such that, as the first of the five pulses of the code signal is being received, the first cam engages the projection on the code bar operating lever associated with the first code bar and rotates it slightly in a counterclockwise direction. Assuming that the received No. 1 pulse is marking, the armature will be operated, and the movement of the code bar operating lever will lift the *sword* and cause its upper right-hand projection to strike the right-hand end of the armature extension. The sword will then rotate in a clockwise direction on its pivot at *A*, so that, when the selector cam clears the code bar operating lever and allows the latter's spring to return it to normal, the sword point will be forced against the left-hand projection of the *T lever*, rotating it in a counterclockwise direction and moving the No. 1 code bar to the right. If the received No. 1 pulse had been spacing, this code bar would have been moved to the left.

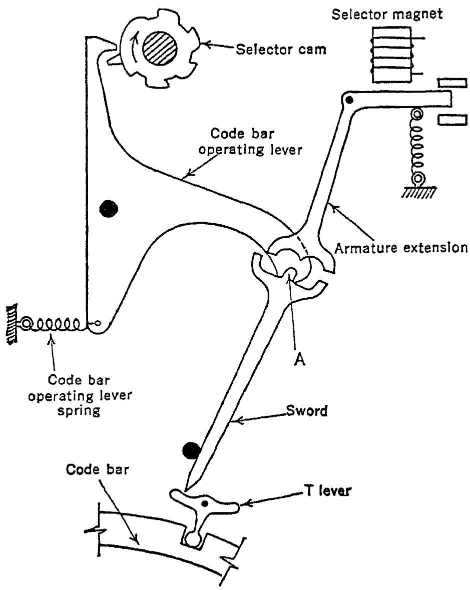


FIG. 16. Teletypewriter Receiving Selector Mechanism, Illustrating Positioning of Code Bars (Courtesy Bell System)

In like manner, the other four code bars are properly positioned and the sixth cam releases a clutch, allowing the printing mechanism to operate, as shown in Fig. 17. The five code bars have been positioned to the right or left in accordance with the five code pulses received, which represent a particular character. The code bar slots will then line up so that the pull bar for that character will center in these and only these slots and will engage the main bail. This bail is then moved forward, causing the pull bar to move forward and the type bar head to be driven against the paper or tape in the teletypewriter.

After completion of one revolution of the rotating cams, the rotating movement stops, owing to action of the stopping mechanism, and the teletypewriter is then ready to receive the pulses for the next character.

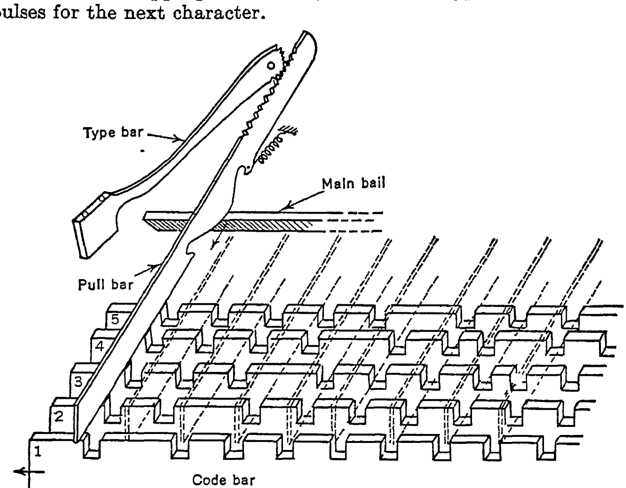


FIG. 17. Teletypewriter Receiving Mechanism, Illustrating Selection of Symbol (Courtesy Bell System)

Power for mechanically operating both the teletypewriter and the teleprinter is generally secured from the 110-volt commercial supply at their location, these machines being equipped with either d-c motors, a-c series motors, or a-c synchronous motors. The last type of motor is available for use with 50- or 60-cycle power and is the preferable type where the power frequency has the usual close regulation. The power consumption is about 115 watts at 115 volts.

The line signaling circuit usually requires about 60 to 70 ma of direct current for magnet and line relay operation, and the remote control circuit (if provided) about 50 ma.

The regenerative repeater may be employed (1) at intermediate points in long commercial or private line teletypewriter circuits, (2) at points where it is desirable to divide a teletypewriter circuit into several sections because of possible distortion from a large number of sending stations connected to the circuit, and (3) at teletypewriter exchanges for switched connections and for furnishing service to subscriber teletypewriter loops having a relatively high transmission coefficient.

Start-stop teletypewriter signals, composed of one start and five signal pulses of equal unit length and one stop pulse of 1.42 units length (teleprinter signals are composed of seven pulses of equal unit length), are sent from the originating motor-driven distributor or other sending device with mechanical precision. However, as these signals progress over a telegraph circuit several telegraph repeater sections long, the distortion accumulates section by section, so that, unless corrected, the distorted signals received at the circuit terminal would result in message errors.

The regenerative repeater located at one or more intermediate points in a long teletypewriter circuit, as required, retimes and reshapes the received distorted signal pulses and retransmits them, as though they were directly from the originating machine.

Figure 18 shows a schematic of the circuit for one two-way regenerative repeater arranged to repeat signals from east to west or west to east, but not in both directions

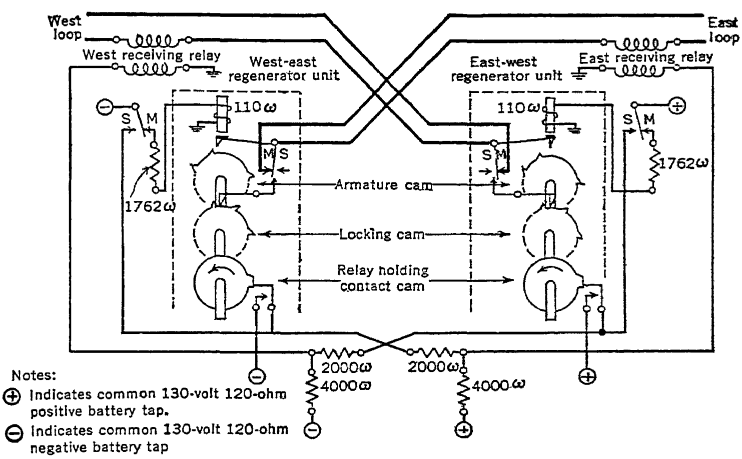


FIG. 18. Schematic Circuit of the Regenerative Repeater, Arranged for Half-duplex Operation (Courtesy Bell System)

simultaneously. For full duplex operation, the circuit requires two one-way regenerative repeaters. The loop circuits are arranged to receive and transmit neutral signals from the terminal duplex telegraph repeaters between which the regenerative repeater is usually connected. Each loop termination contains the line winding of a polar receiving relay and the sending contacts of a regenerator unit in series. Each receiving relay armature controls a local battery circuit to energize the magnet of the regenerator unit through the marking contact. When the receiving relay armature moves to its spacing contact, the receiving relay armature in the opposite loop is locked to marking.

The regenerator unit has both sending and holding contacts, the latter being controlled by a cam in such a way that they are open only when the cam assembly is near the stop position. During retransmission of the signal elements of a complete character, the holding contacts, being closed, apply battery of opposite polarity through a potentiometer to the receiving relay biasing winding of the loop, into which the signals are being retransmitted. Reversal of current in this biasing winding holds its armature to marking, thus preventing repetition of the retransmitted signal back through the regenerative repeater toward the originating terminal.

Since the length of the retransmitted signals is timed mechanically by the locking cam, the outgoing signal length is normally independent of the incoming signal length. As long as that part of the received pulse corresponding in time to the release of the magnet armature is not affected by distortion, the outgoing pulse will be undistorted.

The multiplex system, a development of 1915, functions to divide the line facility time of a given telegraph channel (between two terminal points) among several telegraph circuits. The usual manually operated teleprinter (or teletypewriter) circuit line signal speed is inherently much less than the speed capabilities of a high-grade telegraph trunk. To utilize such capabilities efficiently, the trunk time may be allotted among two, three, four, or more telegraph circuits on a full duplex basis. When four transmitting and four receiving terminals are operating simultaneously at 66 words per minute over one trunk line, this line is handling 528 (8 × 66) words per minute, thus enormously increasing its efficiency over single teleprinter circuit operation. One multiplex channel may be operated in each direction over a duplexed circuit or carrier channel, each of which provides two separate telegraph paths. This system, because of its high efficiency of operation, is generally employed over heavily loaded telegraph trunk lines.

A code is used in which every character consists of 5 equal-length impulses, each of which may be either positive or negative, thus yielding 32 separate combinations. The code as it appears in a perforated transmitting tape is shown in Fig. 13. The black dots, representing perforations in the tape, correspond to marking impulses, and the blank spaces correspond to spacing impulses. One of the 32 possible combinations, 5 positive or spacing impulses in succession, has no character assigned to it, and is transmitted continuously when no messages are being sent. Two combinations, designated "letters" and "figures," are used respectively to cause the printer to print lower-case or upper-case characters, thereby increasing the total number of characters that may be transmitted.

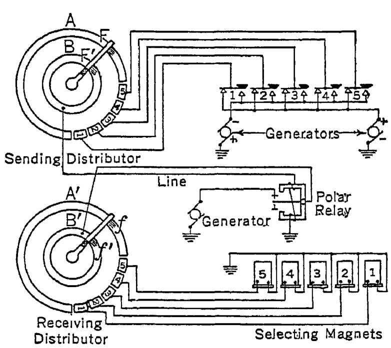


FIG. 19. Schematic Diagram of the Multiplex System

A schematic of this system is shown in Fig. 19. Each of five contacting levers in a transmitter is connected to correspondingly numbered equal-length segments of a distributor, which are successively connected to the line through the rotating brushes F, F' and an unsegmented collector ring B. During the time the brushes are traversing another part of the distributor, the transmitter levers are positioned to correspond with perforations in the transmitting tape. After the brushes have passed over those segments and transmitted the signal combination to the line the tape is advanced to the next set of perforations and the cycle of operations is repeated. Auxiliary or local segments (not shown) in the distributor are used to step the transmitter ahead at the proper point in each revolution of the distributor.

The received signals operate a polarized relay, whose armature applies current from a local generator to the brushes f, f' of a receiving distributor, which comprises a solid collecting ring B' and a segmented ring A'. Five segments of this ring occupying about the same angular position as the sending segments are connected to five correspondingly numbered magnets which control the operation of the printer. The sending and receiving brushes at opposite ends of the line are rotated in nearly exact synchronism, and their angular positions are so adjusted by automatic means that the receiving brush will make contact with one segment at the instant a pulse is received from the similarly positioned sending segment, thus operating the corresponding selecting magnet. Therefore, at each revolution of the brushes the five printer magnets will be operated in sequence to reproduce the identical combination set up on the transmitter levers.

One or more additional sets of transmitters and printers, similarly connected, provide several independent channel transmissions during one complete revolution of the brushes, so that by proper choice of the number of channels the full capacity of a line may be utilized, while the individual channels are operated at speeds that will not exceed the capabilities of the operators or apparatus.

The transmitters are controlled by a perforated sending tape (Fig. 13), prepared by the operator on a perforator which has a keyboard similar to that of a typewriter. Two types of printers are used. One, known as a page printer, is similar to a typewriter and is equipped with automatic means for returning the carriage and for feeding the paper to print the message in page form; the other prints the message in a continuous line on a narrow paper tape, which is gummed on the reverse side to permit of its being readily pasted on message forms.

Distributor brushes are rotated by a synchronous impulse motor of the LaCour type, which is supplied with impulses of constant frequency generated by the contacts of an electrically driven tuning fork. Slight variations in motor speed are compensated for by applying a precise phase correction controlled by intelligence signals received from the distant terminal. Since synchronism is maintained between the sending and receiving terminals, the standard start and stop pulses (of the teleprinter) are not required for the multiplex circuit, but other local control pulses are employed.

Multiplex channels are operated commercially at speeds from 50 to 80 words per minute (one word averages 5 letters and a space, which is equivalent to 15 cycles per word). The number of channels used is determined by the traffic load and type of circuit available. On duplexed land lines a maximum of four channels in each direction is permissible. Three channels can be satisfactorily operated on carrier channels spaced 170 cycles apart, and four multiplex channels may be applied on carrier channels spaced 200 cycles apart. Two to eight channels may be successfully operated on submarine cable, depending on the type of cable and terminal equipment.

The line frequencies for various speeds and number of channels in each direction are given in Table 1.

Table 1. Line Frequencies

5-letter Words per Minute	Line Frequency, cycles per second				
	8 Channels	6 Channels	4 Channels	3 Channels	2 Channels
50	100	75	50	37.5	25
60	120 *	90	60	45	30
70	140 *	105 *	70	52.5	35
80	160 *	120 *	80	60.0	40

* Not used commercially.

Special repeaters are not required for the operation of multiplex on duplexed circuits less than 1000 miles long, as the regular duplex repeaters are usually satisfactory. On longer circuits, or in those which contain more than three duplex repeaters in succession, the signals are likely to be distorted sufficiently to cause frequent errors and require excessive attention to the apparatus. In such circuits a regenerative repeater is usually employed at every third or fourth repeater point.

The varioplex system, a Western Union Telegraph Co. development, is an automatic telegraph system which provides for the connection of up to 40 individual telegraph circuits or subchannels, having variable message loads, over a single high-capacity telegraph trunk circuit, to other individual telegraph circuits or subchannels at one or more distant points. The actual number of subchannels operated over a single varioplex trunk is limited mainly by practical considerations, one important factor being the total message load presented to the trunk at any one time.

Though the cost of this service to the patron is relatively low, the patron has, in effect, a private high-grade telegraph connection to the distant party. The installation for each patron consists of a sending teleprinter and a receiving teleprinter, thus providing simultaneous two-way service. Character counters, connected to the sending legs of each subchannel, determine the number of words sent, for billing purposes.

Messages are transmitted over a given subchannel by operating the keyboard of the sending teleprinter, and messages are received over a separate subchannel and automatically printed by the receiving teleprinter on a page or tape. Any patron may send a message at any time, as desired, and the other patrons in a given varioplex system may do likewise. All the messages are received at the central office in the varioplex terminal equipment and transmitted in sequence over the varioplex trunk to its distant terminal, which automatically distributes each message to the proper distant subchannel, over which it reaches the patron for whom it is intended.

The varioplex terminal equipment may be classified as (1) individual to each subchannel, (2) common to all connections, and (3) part of the varioplex trunk circuit.

Figure 20 shows a schematic of a two-channel (*A* and *B*) varioplex circuit with eight patron offices and teleprinters *TPR* at each terminal of the circuit for one direction of transmission only. The opposite direction is similarly provided. The principal equipment units at sending terminal X for each subchannel include a reperforator *RPF*, tape transmitter *XTR*, sending chain relays *SA* and *SB*, and sending control relay *SCO*. Common equipment units include two banks (five relays per bank) of sending relays *A* and *B* and a segmented distributor sending ring (five segments per varioplex circuit channel).

The principal equipment units at receiving terminal Y include, as common equipment, a segmented distributor receiving ring (matches the segmented sending ring), two banks (five relays per bank) of receiving relays A and B, and two local transmitting devices TA and TB (each an arrangement of vacuum tubes). The individual subchannel equipment includes receiving chain relays RA and RB and receiving control relay RCO.

Five sending and five receiving relays (banks A and A or B and B of Fig. 20) are associated with the five respective segments of each channel. These relays function in such a manner that, if teleprinter characters are being sent by one or more patrons at station X, positive or negative potentials will be applied by each relay to its corresponding segment at station X, causing positive or negative pulses (combination determined by the

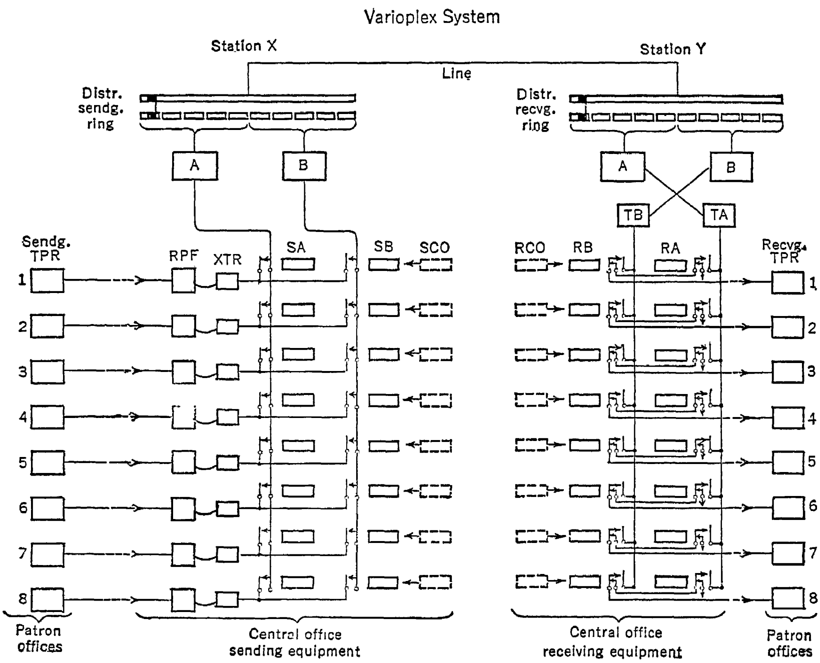


FIG. 20. Schematic of Two-channel (Double-type) Varioplex Circuit (Courtesy Western Union Telegraph Co., Electrical Communication and O. E. Pierson)

character being sent) to be transmitted over the line to station Y, as the sending brushes pass over the segments connecting them, one by one, to the line.

At station Y the distributor brushes rotate in close synchronism with the distributor brushes at station X, and each receiving segment is connected through the line, for a short interval of time, to its corresponding sending segment on the distributor rings. As the pulses are received at station Y, the receiving relays are operated or released, depending on whether the pulses are positive or negative.

The relay RA or RB of a given subchannel will operate in unison with the relay SA or SB of the corresponding sending subchannel and will remain operated during selection of a character by the receiving bank relays A or B. The two relays RA and RB of each subchannel control their subchannel circuit through their contacts. This circuit is closed to a fixed potential, when both relays are in their released positions, but is disconnected from this potential and is connected to device TA or TB, upon operation of RA or RB, respectively. Devices TA and TB are controlled by the contacts of the bank relays A and B, respectively, and by a segmented ring (simplex ring) of the receiving distributor.

Assuming that a given RA relay is in its operated position, pulses from the receiving simplex ring segments actuate the input circuit of the vacuum tube unit TA in accordance with the selected character stored in relay bank A, causing a signal combination to be transmitted over the connected subchannel to the associated teleprinter, which prints the character sent.

No distinction is made with respect to which varioplex channel is used for the transmission of a character from any given sending teleprinter. Successive characters are sent alternately over the two channels, one character being supplied by each subchannel in turn. After all patrons who are sending have transmitted one character, the cycle of operation is repeated. Successive characters from a given subchannel may be transmitted over either varioplex channel.

Relays *SA* and *SB* operate in a cyclic manner, characteristic of the counting chain type of relay circuit. A local ring of the distributor furnishes two pulses per revolution, one pulse for operating each chain relay. Each chain relay is locked in its operated position after being actuated by one of these pulses, and, whenever any chain relay is actuated, any other previously operated relay in the same vertical row (Fig. 20) is released. Not more than one *SA* or *SB* relay is operated at one time. Also, the circuit is such that not more than one relay can be operated in the same horizontal row. If all eight subchannels were sending at the same time, the chain relays would operate progressively upward, starting with *SB* relay 8, then *SA* relay 7, *SB* relay 6, and so on. Thus, in four revolutions of the brush each of the eight sending subchannels would transmit one character. This cycle of operations would be repeated as long as signals were being sent.

The *SCO* and *RCO* relays operate in unison, by means of special signals sent over the line, to remove from the operating chain circuits their associated *SA* and *SB* and *RA* and *RB* chain relays, respectively, when these chain relays are not functioning. This removal does not affect the other chain relays, which are operative. Thus, by controlling the *SCO* and *RCO* relays, so that they remain operative if traffic is available at the transmitter, and inoperative if such traffic is not offered, the main line circuit time is made available only to those subchannels having a simultaneous need for it, and the cycle of operations is speeded up. The *SCO* relay is controlled through a collating arrangement which "reads" the character in the transmitter, causing the sending control relay *SCO* to operate or release, depending on the type of character.

If the line between stations *X* and *Y* is capable of transmitting satisfactory signals at a rate of 800 characters per minute, the two distributors for the two-channel varioplex circuit are adjusted to a speed of 400 rpm. This method may be employed for all high-speed automatic circuits to divide the total traffic capacity of a circuit into smaller and more practical components.

Reperforator switching systems are primarily switching arrangements for use in large message relay centers. These systems not only reduce to a minimum the time required in relaying messages through such centers but also substantially increase the message capacity of trunk lines and facilitate personnel training problems. This equipment is now in operation in a number of the main telegraph centers in the United States, and installations are in progress in other large centers.

The reperforator switching system of recent design consists principally of: (1) receiving equipment, which prints the incoming characters on and perforates them in a tape; (2) crossoffice trunking circuits, by means of which the received message is transferred to a sending position; and (3) the sending equipment, which retransmits the message to its destination or to a second relay center.

Incoming and outgoing transmission as handled by the reperforator-switching system is usually at the rate of 66 words per minute; the rate over the crossoffice trunks is 150 words per minute, thus permitting rapid clearance of messages from the receiving positions. The high rate over the crossoffice trunk is due to the use of a five-wire trunk, permitting the sending of a five-element character at a time. Since the attendant has only two switch operations to perform in order to relay a message through an office having this system, the time required in handling the message is only a matter of seconds. The message is passed through the office entirely on tape, no manual receiving or sending being involved.

If an outgoing channel selected for the transmission of a message is idle at the time of selection, there is no delay at this channel in retransmitting the message. If the channel is busy and no other channel is selected for the message, the tape from the sending reperforator is automatically stored in a specially constructed narrow glass compartment and is fed out through the line transmitter automatically, as soon as the busy channel is available.

Special centers, known as "spillover" and XV centers, of the switching office are provided to handle messages that are abnormally delayed for various reasons, such as circuit trouble, destination office closed, incomplete address, and uncertain routing, or to handle messages of an emergency or special nature. These centers contain switching turrets and printer perforator receiving positions of the same general type, as previously described.

Subcenter switching systems (Western Union Telegraph Co.) are employed, where a number of branch offices and private line patrons are localized in an area some distance from the nearest main switching center. In order to eliminate delays in handling the messages originating from these sources, and to reduce operating costs which would occur if the messages were collected at a local office by messenger or otherwise and thence transferred to a switching center, subcenter switching units, as required, are installed in such areas to extend the local lines automatically direct to the main switching center over a small group of trunks. Usually the number of trunks required is about one-third the number of local telegraph stations served.

In this sytem, the patron's teleprinter sends *outgoing messages* to the main switching center direct, the local line being automatically switched through the subcenter. A direct circuit from the main center to the patron is established for incoming messages by the main-center attendant dialing the patron's line number over an idle subcenter trunk, which is automatically connected to the patron's line through switches at the subcenter.

In Teletypewriter Exchange Service (TWX) in the Bell System, line concentrating units are usually employed for serving areas, such as described above. These units are arranged for automatic switching of two-way message service between a group of subscriber lines at an outlying center and a manual teletypewriter switchboard over a small group of trunks. One such unit has a capacity of 30 lines, which appear on the verticals of three 10 by 10 crossbar switches. The trunks appear on the horizontals of these switches. A 100-line unit is also available.

Private line switching and intercommunicating systems are useful in extensive telegraph leased wire networks, such as are employed to connect a number of widely located stations and offices in large industries or governmental agencies. These systems perform a service in handling telegraph messages somewhat comparable to that which private branch exchanges perform in handling telephone messages. The systems are attractive to large users because of their simplicity to operate and freedom from trouble.

The larger private line switching systems of the Western Union Telegraph Co., receive messages on printer perforators, which feed the perforated tape into a sending tape transmitter. This transmitter is connected to the desired outgoing line by means of a plug-ended cord and jack.

When a message is being received at a given unit, the attendant notes its destination on the tape and inserts the plug of the cord circuit associated with that unit in one of two multipled jacks in which the outgoing line terminates. Two jacks are provided for each line so that, if one of the jacks is in use and a second message is received for that line, the connection may be made for this message in the idle jack and the message will be sent as soon as the first message has been cleared.

Perforator and transmitter units are frequently provided for the patrons, where it is desired to prepare messages in advance by perforating them in a tape and transmitting them at the same time or later.

Fully automatic message switching systems in which the first characters of the code serve to control the switching equipment have been developed and are now in operation.

Intercommunicating systems, as provided by the Bell System for private and governmental users, employ teletypewriter station equipment and teletypewriter switchboards, at which *connections between stations or between a station and a trunk* are established manually by means of cord circuits. The operator is provided with a teletypewriter by means of which incoming calls are answered and outgoing connections are established by typing on a keyboard. The station may have a page type sending-receiving teletypewriter, or a sending-receiving typing reperforator, for printing characters on and perforating them in a tape ($^{11}/_{16}$-in. wide), with which is associated a transmitter-distributor, or other combinations of sending and receiving equipment may be provided.

Teletypewriter Exchange service (TWX), as established throughout the United States by the Bell System for public use, provides both large and small switching centers interconnected by trunk circuits of suitable grade.

Thus, two TWX subscribers, being provided with the necessary equipment, may communicate with each other by written message from one part of the country to another over a vast network of lines and equipment, somewhat similar to that established for nation-wide telephone service.

Transmission limitations, with respect to overall connections, are important factors in furnishing a satisfactory general teletypewriter exchange service, as discussed in article 16.

Teletypesetting, a process of automatically setting type, is accomplished by perforating in a tape the copy material to be set in type, and then feeding the perforated tape into a transmitting device which actuates a typesetting machine. The tape may be perforated locally or by a perforator receiving automatic signals of the proper code over a telegraph circuit from a distant point. The six-unit code is used in the transmission of the characters, and special teletypewriter or teleprinter equipment arranged to send and receive this code is employed, providing page copy for checking purposes.

8. ALTERNATING-CURRENT TELEGRAPH SYSTEMS

Alternating-current telegraph systems employ both wire and radio channels for the transmission of telegraph signals.

Voice-frequency carrier telegraph systems operate within the lowest carrier frequency (voice) range of about 250 to 3150 cycles.

One such system, now in common use, provides up to 18 two-way telegraph channels having carrier frequencies from 255 to 3145 cycles, spaced 170 cycles apart. This particular system operates on a four-wire line basis over various types of facilities, such as loaded four-wire cable circuits, open-wire physical circuits on a four-wire basis (noise conditions permitting), and the different types of carrier telephone channels under suitable conditions.

Figure 21 shows the principal elements of a telegraph channel from the sending to the receiving telegraph terminal for the particular eighteen-channel system referred to above. At the sending terminal, the marking signal closes the line circuit at the step-up transformer and the carrier current is transmitted over the line to the receiving terminal, where it is amplified and detected, and finally operates the receiving relay to marking (−). For

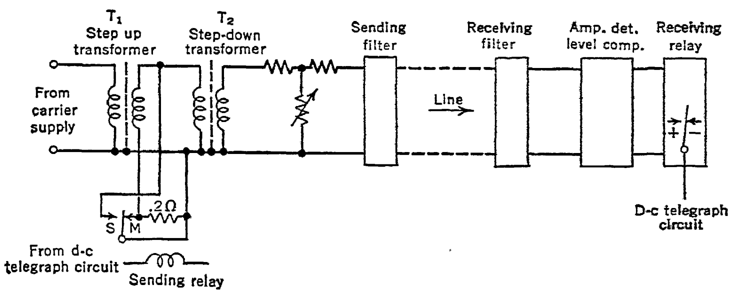


FIG. 21. Principal Elements of a Carrier Telegraph Channel (Courtesy Bell System)

the spacing signal, the input of transformer T_2 is short-circuited at the step-up transformer, and no carrier current reaches the line. The receiving relay is operated to spacing (+), when no current is being received from the line. The 0.2-megohm resistance inserted in series with the line winding of transformer T_1 assists in reducing the carrier to the line and lessening the load on the carrier supply during spacing in this system. The send and receive filters pass only the carrier frequency assigned to the particular channel with which they are associated.

The *telegraph level* at a given point on the circuit is the power at that point due to a single steadily marking channel and is expressed in decibels referred to 1 milliwatt (dbm). If frequency distortion is present, the level refers to a channel operating at or near 1000 cycles, which is the *nominal* telegraph level. The telegraph levels on the line section of a circuit, in general, must be high enough to override line interference and low enough to prevent modulation and crowding, which introduce objectionable interference in other systems. The sending and receiving levels in the same telegraph system should be such that cross-induction will not occur between the sending and receiving branches in the terminal cabling and equipment.

Specific telegraph level (STL) of a circuit used alternately for telephone service is numerically equal to the power of one telegraph channel (in dbm) at a point of zero telephone transmission level. At present in the Bell System, this level is −16 db for type C carrier (except in certain cases) and −21 db for other telephone facilities. The proper selection of the *STL* in any given case is highly important in the application of telegraph to telephone circuits, because of the dependency of the satisfactory operation of both services, when related, on this factor.

The telegraph signaling speed that can be obtained on a carrier channel depends on the band width or frequency range of the channel. With wider bands, higher signaling speeds per channel are possible but fewer channels are obtained. When the carrier channels are spaced 170 cycles apart, each channel will allow telegraph signaling at speeds of 35 to 40 cycles per second, which permits of the operation of a three-channel multiplex working at 50 words per minute per multiplex channel. Thus, a twelve-channel carrier system operated in this way would have a total capacity of 1800 words per minute in each direction simultaneously. With 300-cycle spacing between carrier channels, telegraph speeds as high as 75 cycles per second may be obtained on each channel, which permits the operation of a four-channel multiplex at 75 words per minute per multiplex channel. An eight-channel system of this type has a capacity of 2400 words per minute in each direction simultaneously.

Voice-frequency carrier systems are usually operated over four-wire circuits, one pair being used for the channels working in one direction and the remaining pair for channels working in the opposite direction. They may be operated on two-wire circuits, however, by transmitting the lower half of the carrier frequencies in one direction and the upper half in the opposite direction in a manner similar to the high-frequency carrier, which is described in this article.

Standard voice or carrier telephone repeaters are employed for voice-frequency carrier telegraph systems assigned to voice or carrier telephone facilities. These repeaters take the same spacing as would be employed if the facilities were assigned to telephone circuits.

Carrier supply for the different carrier telegraph frequencies in the type of system shown in Fig. 21 is now generally furnished by vacuum-tube oscillators, although some motor-driven multifrequency generator sets are still in service. One vacuum-tube oscillator unit has a capacity for as many as 50 carrier telegraph channels normally but for a short time may supply a much greater number of channels. Harmonic control is provided for the vacuum-tube oscillator supply to limit high peak line currents, which would tend to overload the telephone repeaters and, in carrier systems, cause interchannel interference.

Voice-frequency carrier transmission is sometimes employed between a teletypewriter office and an outlying teletypewriter station, particularly where a d-c telegraph channel cannot be satisfactorily provided.

In this system, a carrier frequency of 690 cycles transmitting from the office and 1640 cycles transmitting from the station is used. Carrier current is transmitted for spacing and no current is transmitted for marking signals. The break relay at the office can thus be omitted, since the sending relay opens the receiving circuit during transmission of a spacing signal, which holds the receiving relay on its marking contact. Also carrier current is not sent over the line from either terminal while the receiving circuit at that terminal is connected to the line.

Filters are required at both terminals, in order to prevent echo current effects from reaching the detectors and causing false signals.

The series resistance across the low-pass filter at the station provides for spacing signal feedback to the station amplifier detector circuit and permits obtaining a home copy of the outgoing-station message.

The operation of this carrier system does not affect the usual telephone circuit transmission over the line and is not materially affected by reasonable amounts of line leakage or by earth potentials, crossfire, or power induction in the circuit.

High-frequency carrier telegraph systems for open-wire application have been employed by the larger communication companies for a number of years. The ten-channel open-wire systems of the Bell companies are no longer standard, owing to the improvements and economies secured from the eighteen-channel voice-frequency system described above.

The Western Union Telegraph Co. has developed a series of carrier telegraph systems specially designed to fill the requirements of the domestic telegraph system. A number of these systems are in operation (1947), and additional systems are being installed rapidly in a modernization program in which carrier operation will replace d-c telegraph operation for trunk circuits and will substantially reduce the number of wires and pole lines required for the telegraph service. These systems include (1) a portable type for establishing short temporary or emergency channels, (2) a low-cost four-channel system for distances not greater than two repeater sections, and (3) four types of multichannel, long-distance trunk systems. The basic, standard unit of the trunk carrier systems is a 3000-cycle voice-frequency band extending from 300 to 3300 cycles. The four trunk systems are (1) the 7.5-kc type E with one voice-frequency band in each direction, (2) the 15-kc type F with two bands in each direction, (3) the 30-kc type G with four bands in each direction, and (4) the type WN with 32 bands in each direction and requiring a transmission band of 150 kc in each direction. Types E, F, and G are open-wire systems designed for two-wire operation and with frequency allocations (for 300-cycle channels) as shown in Fig. 22. The type WN is designed for use with a two-way microwave radio relay circuit.

In multiband systems, only one band can be transmitted in its original frequency position (300 to 3300 cycles); all other bands must be transferred or translated to separate positions in the available frequency spectrum of the transmission medium. This can be accomplished by modulating each of the bands requiring translation by a separate secondary carrier of appropriate frequency. Such a method is wasteful of the frequency spectrum because of the increasing inefficiency (in absolute band width) of selective filters as the frequency is increased. The Western Union systems utilize the spectrum more efficiently by employing a plural or tandem method of modulation in which the voice-frequency bands are translated in groups, thus requiring only broad group filters at the assigned position in the frequency spectrum. The secondary carrier or translation frequencies are supplied by amplified harmonics of a high-stability, low-frequency oscillator.

Although these carrier systems are designed primarily for operation over Western Union facilities, they are readily adapted to operation on any wire or radio transmission system that provides a suitable transmission band. Furthermore, the individual voice-frequency bands can be repeated or patched at will between the four types of Western Union multiband systems, Bell System facilities, and equivalent facilities of other companies.

The voice-frequency bands may be utilized for high-speed facsimile telegraph transmission or for operation as a telephone circuit, but they are ordinarily channelized for

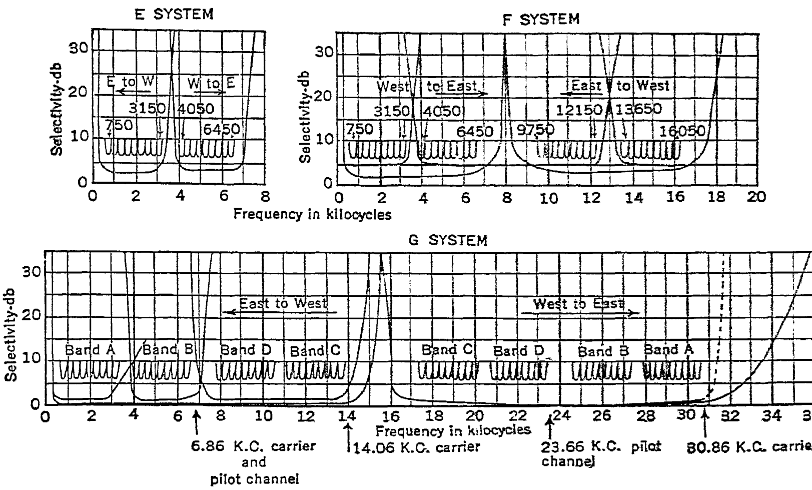


FIG. 22. Frequency Allocations for Three Types of Carrier Telegraph Systems (Courtesy Western Union Telegraph Co., A.I.E.E., F. E. D'Humy, and P. J. Howe)

printer or other code telegraph operation. Two types of channelization are employed, providing (1) nine-wide-band channels spaced 300 cycles apart and suitable for four-channel multiplex operation at speeds up to 75 cycles per second, and (2) eighteen narrow-band channels spaced 150 cycles apart and suitable for teleprinter or two-channel multiplex operation. Present trends are away from the relative mechanical complexity of multiplex telegraphy and toward teleprinter operation; consequently, narrow-band channels only are now being installed.

The following description deals with narrow-band channels, but it also applies generally to the basically similar wide-band equipment. The sending terminal of each channel includes a vacuum-tube oscillator, which supplies the channel carrier frequency. The d-c telegraph signals are impressed or superposed on the carrier by modulation. The carrier telegraph systems of other companies use amplitude modulation, as did the early Western Union systems. For amplitude-modulated (a-m) operation, the channel carrier frequency is transmitted for a marking signal and interrupted for a spacing signal. Western Union has introduced, and now uses exclusively, frequency modulation for carrier telegraph transmission. In f-m transmission, the frequency of the channel oscillator is controlled by the d-c telegraph signals being transmitted. Marking and spacing signals cause the frequency to swing down and up, respectively, 35 cycles from midchannel frequency. The transitions between marking and spacing frequencies are not instantaneous but are very abrupt, with the frequency increasing and decreasing smoothly between the lower and upper limits. This action results in true frequency modulation, and the modulated carrier current appears at the output of the channel transmitting filter constant in amplitude but varying in frequency.

At the distant corresponding channel terminal, the carrier is selected by the receiving filter and amplified, after which it passes through an amplifier limiter, a frequency discriminator, and a rectifier circuit, which restores the marking and spacing frequencies to their original d-c form. The vacuum-tube limiter circuit maintains a constant-amplitude output for received levels in a range between -50 and $+10$ dbm.

In comparison with a-m transmission, f-m transmission is definitely less susceptible to interference, and the received signals are immune to the influence of level variations encountered in transmission.

The original f-m channel terminals contained transmitting and receiving relays. In the latest type of narrow-band equipment, the relays have been eliminated and operation of the channel terminals is completely electronic. This change, together with other simplifications, has effected a considerable reduction in size and cost of equipment.

Transmission losses in the open-wire systems between sending and receiving terminals are compensated for by vacuum-tube repeaters, spaced at suitable intervals along the line. Average wet-weather losses are limited to about 25 db between repeaters. Reflection losses at junctions of open wire and incidental intermediate or terminal cables are limited by impedance-matching networks. Loading of short cables to reduce attenuation losses is no longer favored. Networks are provided at repeaters and terminals to equalize the losses at different frequencies, so that the attenuation is substantially uniform over the carrier band.

Telegraph circuits can be operated across the United States entirely by carrier without requiring a regenerative repeater at any point. The type G system provides, in each direction, on a single pair of wires 72 teleprinter circuits or 144 multiplex channels having a total capacity of 4750 and 9500 words per minute, respectively, in each direction. The other systems provide capacities proportionate to the number of voice-frequency bands employed. The radio relay system with its 576 carrier channels handles 38,000 words per minute each way.

9. FACSIMILE SYSTEM

Facsimile, as defined and discussed in detail elsewhere in this handbook, is a process whereby such objects as a picture or a sheet of paper containing printing or writing are electrically scanned, and the electrical currents thus generated are transmitted by wire or radio to a receiving device which reproduces, as a print or on a specially prepared paper, the original picture (in black and white) or the printing or writing. The following brief discussion applies to facsimile as used in telegraph operation only.

Facsimile transmission of messages, as employed by the Western Union Telegraph Co., is known as *telefax* service or transmission. This type of service was considered desirable by the telegraph companies for many years, but the recording systems used, employing photographic, chemical, or other processes, were too costly or slow for the general handling of telegrams.

In recent years, the Western Union Telegraph Co. developed for its use a facsimile recording paper, with the trademark *Teledeltos.* This paper has a conducting coating of a light gray color which is marked in black by the passage of an electric current through it. The paper is used dry, requires no processing, is not affected by light or ordinary moisture conditions, and produces immediately a clear-cut permanent record. Simple Telefax recording equipment receives the incoming amplified facsimile currents and applies them directly to the paper by means of a stylus riding on the surface of the paper.

The Telefax equipment thus far developed is of several different types and has been designed primarily to provide maximum convenience for the general public. However, trunk-line Telefax is being used to some extent between large centers, as between Chicago and New York. These trunk circuits have been used mainly for the transmission of drawings, sketches, copy for publication, editorial corrections, and commercial messages. For good-quality Telefax transmission, a frequency band width of about 2500 cycles is required. The line loss at the maximum frequency used should not exceed about 25 db.

10. MISCELLANEOUS TRANSMITTING AND SIGNALING SYSTEMS

The ticker system is designed to furnish stock, bond, and grain quotations and pertinent news items to brokerage, investment, and private offices during trading hours.

The ticker, which does not include a transmitting keyboard, is otherwise quite similar in principle to the start-stop printer, the chief points of difference being in the form of type wheel used and the method employed in shifting from letters to figures. An eight-unit code is used. The first impulse of each group starts the printer; the second impulse determines by its polarity whether letters or figures are to be printed; the succeeding five pulses determine the particular character to be printed; and the eighth pulse is the stop pulse, which allows the distributing mechanism to come to rest in preparation for the next succeeding signal group. The tickers are controlled by a single transmitter at the central distributing point. Normal operating speed is from 450 to 500 printed characters per minute.

The teleautograph system is employed principally by banks, railroads, department stores, and similar businesses. This system provides a means by which messages, written in longhand at one station, may be reproduced simultaneously at one or more stations at various locations without material distortion of the original characters.

The principle on which this system operates is shown in Fig. 23. The transmitter consists of a stylus, which is mechanically connected, through two sets of levers and appropriate swivel joints, to the contact arms of two variable rheostats in such a way that the horizontal and vertical components of the stylus movements are translated into corresponding current variations in two lines connecting the receiver. At the receiver, the variations in the line currents produce similar movements in two coils or "buckets" within a magnetic field. The movements of these coils are communicated through a system of levers to a writing pen which reproduces the movements of the sending stylus.

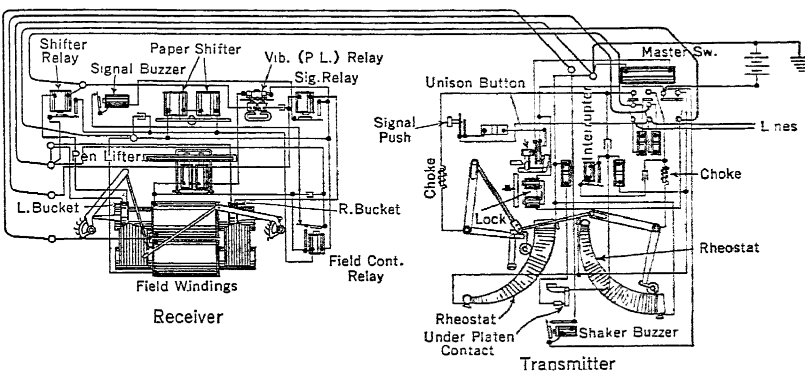


FIG. 23. Schematic Diagram of the Telautograph System

The pen is lifted from the paper and the paper is shifted to provide fresh writing surface by magnets, which are controlled by superposed alternating current supplied by a buzzer located in the transmitter.

Two grounded circuits are required, one to transmit the current variations representing each of the two components of the stylus motion. As many as 100 receivers, arranged to be controlled by one or more transmitters, may be operated in multiple on a pair of wires. The speed of operation is determined entirely by the rapidity with which the operator can write. The line potential is usually 120 volts with one side grounded, and current depends upon the resistances of transmitters and receivers. One type of equipment, which normally operates on 100 ma, is adapted for use on wires not exceeding 700 ohms resistance each. Another type requires only 60-ma line current and will operate well on lines up to 1200 ohms resistance, although in a few extreme cases operation has been maintained on lines having resistances as high as 1800 ohms.

Messenger call circuits are employed for summoning telegraph messengers to pick up messages at various locations and deliver them to the telegraph office for sending.

Signals are transmitted by a call box, which consists of a spring-driven clockwork mechanism arranged to turn a pair of notched contact wheels through one complete revolution each time the box is operated by turning and releasing a winding key, as shown in Fig. 24. The contact wheels are arranged both to open and ground the line when transmitting signals, and to restore the line to its normal closed ungrounded condition upon coming to rest. The call boxes are connected in series with a line having both ends terminated at the central office in relays which operate a buzzer and a register for recording the signals on a paper tape. Switches S_1, S_2, and S_3 are provided to change the line, battery, and register connections to permit of receiving signals from the call boxes even during times when the line circuit may be accidentally open or grounded or both. Only the simultaneous occurrence of a fault on both sides of a call box or group of boxes will prevent the transmission and correct reception of signals. The normal line current in call circuits is 50 ma, and, for satisfactory service, not more than 50 call boxes should be included in any one circuit. Signaling speed usually does not exceed 4 impulses or 2 cycles per second.

Though many technical advances have been made in the art of telegraphy, the messenger call box, of which there are over 300,000 in the United States in offices, hotels, and other public places, still plays an important part in the collection of commercial telegraph messages.

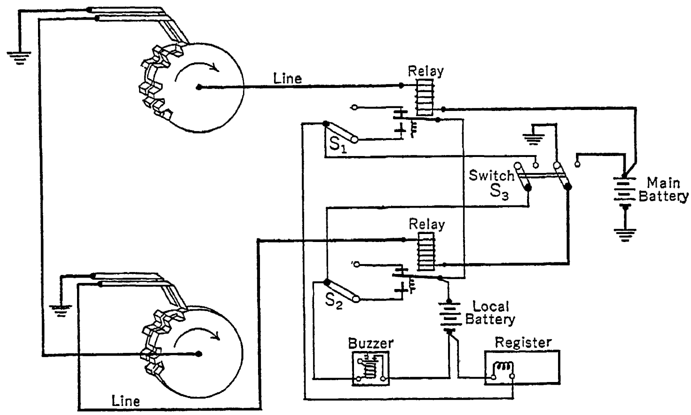


FIG. 24. Schematic Diagram of a Messenger Call Circuit

Clock circuits are employed in furnishing telegraph time service. The clocks at subscriber premises are driven by springs which are wound periodically, usually once each hour, by a small electric motor operating on dry batteries. The minute and second hands are arranged so that they may be moved to their 12 and 60 positions, respectively, by the operation of a synchronizing magnet. The synchronizing magnets of the clocks are connected in series in a grounded circuit terminating at the central office on the contacts of a transmitting relay or a synchronizing machine which sends a synchronizing impulse 1 sec long once every hour, in response to the operation of contacts of a *master clock*. The transmitting circuit is arranged to give an audible signal if the synchronizing impulse fails to be transmitted owing to line failure. The clocks mechanically lock their synchronizing mechanisms in the inoperative position except for two or three minutes immediately preceding and following the time at which the synchronizing impulse is to be received, so as to protect the clocks from being set to a false position by accidental crosses between the line and power circuits. The normal line current is 250 ma, and approximately 60 clocks may be operated on one circuit.

Naval Observatory time signals are regularly distributed to all parts of the United States by the Western Union Telegraph Co. over about 200,000 miles of wire network. These signals provide the means for maintaining some 2000 master clocks, so that they continuously indicate time to a practical degree of accuracy.

Railroad communication systems employ the latest types of telegraph as well as telephone and radio facilities for the control of train movements and the general business of the railroads. The telegraph facilities used are, in general, similar to those previously discussed in this section and will not be considered further here. Special arrangements of telegraph facilities are employed by the railroads to meet their special needs.

SUBMARINE CABLE TELEGRAPHY

By John D. Taylor

Submarine cables interconnect all the earth's continents for the transmission of telegraph messages.

One large telegraph company has more than 30,000 miles of ocean cable, some of it lying at a depth of nearly 3 miles. This company laid its first cable in 1873 and its latest cable in 1928. The North Atlantic is spanned by 14 cables, and 6 cables extend between North America and the Azores, where connections are made with Europe and, via the Cape Verde Islands, with Africa and South America. There are other cables between the United States and Mexico, South America, and the West Indies. Three cables span the Pacific Ocean.

11. CABLE DATA

The first transoceanic cables laid were of the non-loaded type. It was not until 1924 that the first loaded cable was placed, and this cable connected New York with Horta, Azores Islands.

Non-loaded cables of the deep-sea type usually consist of a single copper conductor, which may be solid, stranded, or a combination of both, to provide greater flexibility, as shown in Fig. 1. This conductor is encased in gutta-percha * insulation, covered with servings of jute yarn over which steel armor wire sheathing is placed to provide mechanical strength and protection. The armor wires are covered with inverse layers of jute over which an outer covering of compound is applied. The overall diameter of deep-sea cable varies from about 3/4 in. to 1 1/2 in., depending on the size of conductor and the construction employed.

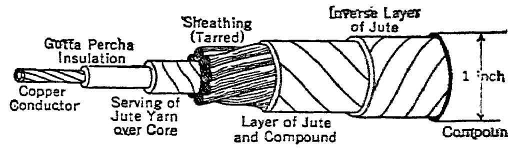


Fig. 1. Non-loaded Submarine Cable Construction

Table 1 shows certain properties of non-loaded cables now in operation.

Table 1. Certain Properties of Non-loaded Deep-sea Cables

Weight, pounds per nautical mile		Diameter of Conductor, mils	Diameter over Gutta-percha, mils	Resistance, ohms per nautical mile at 75 deg fahr	Capacitance, microfarads per nautical mile
Copper	Gutta-percha				
70	120	70	252	16.90	0.272
107	120	86	258	11.05	0.316
107	166	86	298	11.05	0.280
130	130	95	270	9.10	0.334
140	140	99	280	8.45	0.335
160	150	106	291	7.40	0.345
180	160	112	302	6.58	0.351
200	180	114	318	5.92	0.339
225	225	124	354	5.25	0.332
275	225	138	360	4.30	0.363
350	300	151	412	3.38	0.347
500	315	180	432	2.37	0.398
650	400	203	487	1.82	0.398
700	360	211	470	1.69	0.435

The *shore ends of ocean cables* are usually of the twin-conductor type, a cross-section of which is shown in Fig. 2. The second conductor in the end cable is used to extend the circuit ground terminations out into deep water for the purpose of minimizing extra-

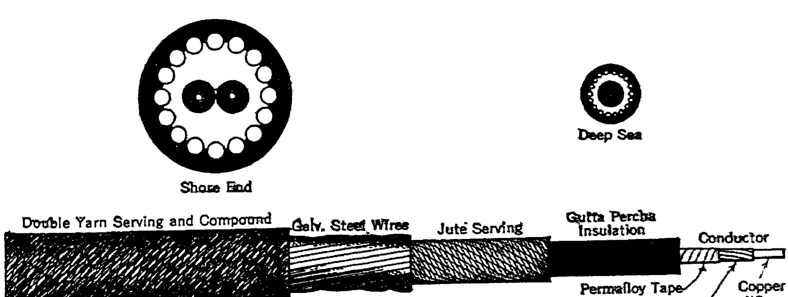


Fig. 2. Loaded Deep-sea Cable Construction

neous ground disturbances from power circuits and natural causes. Since the end sections of cable usually rest in shallow water and are subject to severe water action and other

* Deproteinized rubber has been employed as a substitute for gutta-percha, and a synthetic insulation of polyethylene is being experimented with.

disturbances near shore, the construction is much heavier than for the deep-sea sections, both in armoring and jute layers, so that the overall diameter may be as large as 4 in.

Loaded cables, in their make-up, closely resemble the non-loaded cables just described, the principal difference being that the copper conductor or conductors (for twin conductor) are spirally wrapped with a thin, high-permeability tape of Permalloy (alloy of nickel and iron) (see Fig. 2). This tape uniformly and continuously loads the copper conductors and results in material signal transmission improvements (over the non-loaded cable), such as lower and more uniform attenuation with frequency and less distortion.

Table 2 shows certain properties of loaded cable now in service.

Table 2. Certain Properties of Loaded Deep-sea Cables

Weight, pounds per nautical mile			Diameter in Mils-over			Resistance, ohms per nautical mile at 75 deg fahr	Capacitance, microfarads per nautical mile	Inductance, millihenries per nautical mile
Copper	Gutta-percha	Loading Material	Con-ductor	Load-ing	Gutta-percha			
573	387	72	180	192	480	2.09	0.370	63
517	355	61	171	182	430	2.31	0.375	86
255	252	43	121	132	360	4.65	0.318	140
277 *	258	73	126	148	375	4.28	0.340	170
605 *	370	104	182	202	....	1.97	0.393	118

* Approximate values.

12. OPERATION

Long non-loaded submarine cables (the longest being about 3500 nautical miles) have high values of resistance and capacitance, which attenuate and distort the telegraph signals to such an extent that, until recently, the usual methods of land line operation could not be employed.

Early methods of transmission employed a modified form of bridge duplex with artificial line, giving duplex operation, and electromechanical types of receiving equipment.

Signals were transmitted in the cable Morse Code from a perforated tape by a transmitter and a group of associated relays, arranged to apply positive or negative battery or ground to the apex of the bridge circuit of the duplex set in accordance with the code.

The receiving instrument first used on transatlantic cables was a moving-coil-type mirror galvanometer, connected as such in the bridge duplex circuit. Because of its sensitivity and favorable signal-shaping characteristics, this instrument was used for many years, until replaced by the siphon recorder, which recorded on a moving paper tape in ink the variations in magnitude of the incoming signals. The recorder was later supplemented by (1) various types of magnifiers, which amplified the received signals before the signals reached the recorder, and (2) sensitive cable relays, such as the drum and gold-wire types, which permitted dispensing with manual relay operation at repeater stations.

These developments, together with the application of regenerative repeaters, extending to the period immediately after World War I, made possible greater signaling speeds, but with recorder operation it was still necessary to transcribe messages manually from the recorder tape at the receiving terminal. Satisfactory operation required that the unbalance current in the receiving arm of the bridge should not exceed one-sixth the value of the received signaling currents, thus necessitating the maintaining of a duplex balance between the cable and artificial line within about $^1/_{100}$ per cent.

Because of the high cost of submarine cables, intensive study and experimental work have been continuous, new techniques and devices being sought for increasing the efficiency of these cables. The electromechanical receiving equipment was necessarily fragile to respond to weak signals, and the gain of the magnifiers was relatively low as compared to electronic amplifiers. Attempts in 1918–1919, to apply vacuum-tube signal-shaping amplifiers to improve signal reception did not result favorably, mainly because of (1) high level disturbances existing at that time on duplexed non-loaded cables, due to interference and duplex unbalances, and (2) the fact that suitable electrical networks, equivalent to or better than the mechanically tuned moving coil, were not available.

With the laying of the first loaded oceanic cable in 1924, higher signaling cable speeds were possible, limited by recorder operation and other terminal equipment. Concentrated effort toward improving this equipment resulted in the development of a signal-shaping amplifier and a multiplex printer system suitable for high-speed loaded-cable operation. This equipment was installed on the first and subsequent loaded cables, resulting in raising the message capacity of these cables as much as three to eight times over that of the older, non-loaded systems. These loaded systems are still giving satisfactory service.

The substantial gains made in cable message transmission through development of the loaded-cable system called forth increased effort toward bettering the non-loaded cable performance. The results so far attained have been successfully applied quite extensively in the North Atlantic and in the Alaska communications cable systems, having been accelerated by the needs of World War II.

The improvement program in non-loaded submarine cable operation has included (1) conversion from cable code recorder operation to five-unit code printer operation, (2) replacement of magnifiers with vacuum-tube signal-shaping amplifiers which permit the use of rugged land-line-type polar relays, and (3) improvements in duplex artificial line networks and in the technique of balancing. Thus, the advantages of more nearly automatic operation, increased circuit speeds, and reduced maintenance at cable stations have been secured.

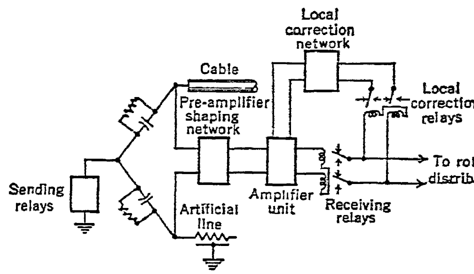


FIG. 3. Typical Arrangement of Terminal Units for Duplexed Non-loaded Cable (Courtesy Western Union Telegraph Co., A.I.E.E., and C. H. Cramer)

The printer system standardized for non-loaded cables is fundamentally similar to the loaded-cable system but more nearly resembles the land line multiplex. It makes use of the tape transmitter, rotary distributor, synchronizing mechanisms, and printers, thus providing for integrated operation with land line systems. It differs from these latter systems in that it is applied to single as well as multichannel operation, whereas the land line system requires start-stop seven-unit code printers for single-channel use; and on long cables the printer signals are transmitted at a speed such that pulses of unit or dot length are received at very small amplitude and, in effect, are considered as absent.

The receiving networks are adjusted to respond to signals two or more units long, so that, from the standpoint of signal reception, the fundamental received frequency is one-half the transmitted dot frequency. The receiving relay operates on a three-position basis, remaining at the zero position for dot signals, which are reinserted synchronously by the receiving rotary distributor. The attenuated-dot method of transmission, though appearing to permit doubling the cable output as compared with normal multiplex transmission, only approaches such output as a limit. Actually, the increase in letters-per-minute circuit speed is only about 80 per cent over that attainable with normal transmission, because, with attenuated-dot transmission, the received signals are more susceptible to interference and more difficult to shape. With this method, the five-unit code used in land line multiplex functions as a 2.5-unit code, giving a net gain over the 3.7-unit recorder Morse code.

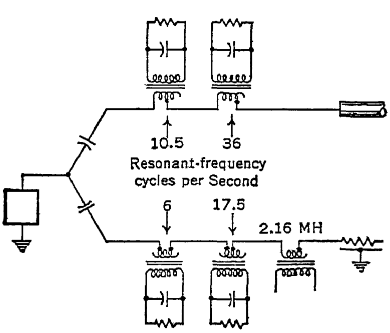


FIG. 4. Typical Arrangement of Resonant Balance Networks (Courtesy Western Union Telegraph Co., A.I.E.E., and C. H. Cramer)

The cable printer system is flexible in that individual channels may be terminated, extended, or combined with other channels to satisfy traffic requirements and the transmission speeds of the available circuits. Channel efficiency may be increased by the application of land line automatic systems.

Figure 3 shows a typical arrangement of important terminal units for duplexed non-loaded cable, including a pre-amplifier shaping network, an amplifier unit, receiving relays, and a local correction network. Resonant balance networks, as shown in Fig. 4, are also provided in a balanced arrangement between the series condensers and the bridge points of the receiving circuit.

The pre-amplifier shaping network, shown diagrammatically in Fig. 5, has been designed to meet certain requirements which permit satisfactory printer operation on long non-loaded cables at the highest practicable speeds, the more important being:

1. A higher standard of receiving accuracy, continuity, and reliability of operation than with recorder operation.
2. The ability to equalize or restore the received frequency components of the signals to an approximation of the original amplitude and phase relationships.
3. Passing only the narrowest band of frequencies, consistent with avoiding undue characteristic distortion and with limiting the effects of extraneous interference and duplex unbalance voltages. Direct currents and alternating currents of near-zero frequency must be rejected, in order to avoid effects from earth currents, particularly as a result of magnetic storms.
4. A shaping network electrically symmetrical with respect to the duplex bridge or electrically isolated from the bridge.
5. Network elements designed for a wide range of adjustment.
6. Amplifier gain and output sufficient to operate rugged polar relays, similar to those used in land line systems.

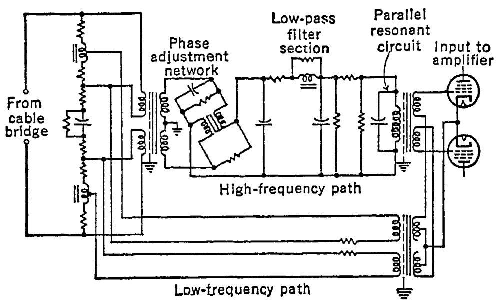


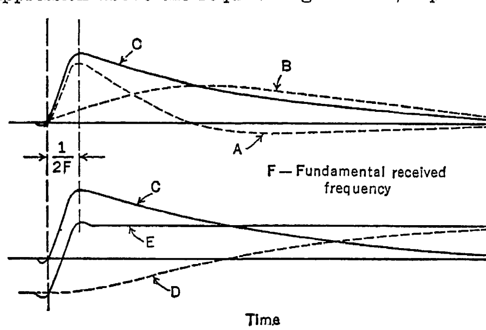
Fig. 5. Pre-amplifier Signal-shaping Network (Courtesy Western Union Telegraph Co., A.I.E.E., and C. H. Cramer)

This pre-amplifier network, when properly adjusted, passes a band of frequencies, including those components required for good signal shape, properly proportioned and phased, and largely or almost completely suppresses those frequencies below the fundamental received frequency. The first branch of the network, being symmetrical and directly across the duplex bridge, is tuned to about 1.5 times the received frequency. The remaining elements of the network, being isolated from the bridge by shielded transformers, are arranged in two paths which combine at the amplifier input.

The lower frequencies pass to the amplifier over the low-frequency path with little, if any, further shaping, while the higher frequencies pass to the amplifier over the high-frequency path, in which they are further shaped by a bridge-type phase-adjusting network, a low-pass filter for added suppression above the required signal band, a parallel resonant circuit tuned to 1.5 times the received frequency, and suitable resistance controls. The phase-adjustment network provides the required wave-front steepness with less higher-frequency components, thus permitting further discrimination against unwanted frequencies.

The amplifier unit is of the three-stage resistance-capacitance coupled, pushpull type with two stages of voltage amplification. Frequencies below those received from the shaping network are suppressed. The maximum voltage gains and available overall gains are 83 db and 103 db, respectively. The amplifier output is adequate for operating the latest land-line-type polar relays.

Two standard two-position relays function in unison as a three-position relay, in accordance with usual cable practice, but the circuit is readily convertible, if desired, to receive two-current signals.



A. Component through high-frequency path of Fig. 5.
B. Component through low-frequency path of Fig. 5.
C. $A + B$.
D. Component supplied by local correction network.
E. Complete signal, $C + D$.

Fig. 6. Received Signal Resulting from Transmission of Long Signal over Non-loaded Submarine Cable (Courtesy Western Union Telegraph Co., A.I.E.E., and C. H. Cramer)

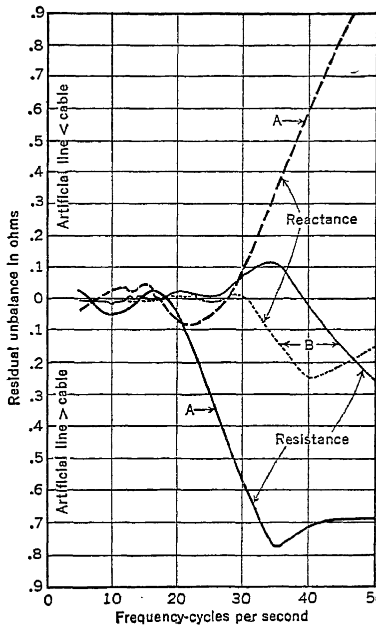
Figure 6 shows the shapes of the component signals in the formative stages and the shape of the complete signal at the amplifier input. Curve *C* shows the signal shape as it leaves the pre-amplifier network, at which point signals of the fundamental received frequency are fully shaped but longer signals would be badly distorted because of lack of low-frequency components. The local correction network (Fig. 3) functions to restore these components under control of the receiving relays. The shaped local correction voltages, curve D, are added to the received signal in the grid circuit of the output stage of the amplifier, resulting in the fully shaped and complete signal, curve E.

For loaded-cable systems, essentially all signal shaping occurs in the pre-amplifier and interstage networks, but there is some deficiency in the very low-frequency components. The method of shaping employed with the non-loaded cable systems not only affords greater immunity from low-frequency disturbances but simplifies amplifier design and stability and eliminates slow transients (wandering zeros).

Duplex operation of non-loaded cables, giving greater total message capacity, is normally used. With the advent of modern amplifiers, extraneous interference and duplex unbalance levels limit signaling speeds. If unbalance is governing, maximum capacity is obtained by using unequal speeds in the two directions of transmission, but duplex unbalance is at present less of a factor, owing to improved artificial lines and balancing methods.

The basic ocean cable artificial line is still about the same as it was at the beginning of duplex operation. The lumped series resistances and shunt capacitances, simulating corresponding cable constants, have been subdivided and arranged for greater flexibility of adjustment. Modern artificial lines of American design are subdivided so that the lumped values of resistance and capacity increase progressively as their distance from the head (point nearest the apex) of the artificial line increases.

The cable circuit parameters include inductance and effective resistance, which vary with frequency because of the earth-return path characteristics, and are known as the *sea-return impedance*. These factors, though small, are important in high-accuracy balancing of the near end of the cable. By using tapered resistance values in series with the shunt capacitance elements of the artificial line sections, the sea-return impedance can be simulated over a relatively wide frequency band and the propagation constant of the sections to which these resistances are added is not materially changed.



A. Before insertion of networks of Fig. 4.
B. After insertion of networks of Fig. 4.

FIG. 7. Frequency Characteristic of Unbalance on a Transatlantic Cable (Courtesy Western Union Telegraph Co., A.I.E.E., and C. H. Cramer)

However, in order to meet balance requirements, further refinements in duplex balancing are necessary. Slow reversals are transmitted, and observations of the residual unbalance transient are made with a cathode-ray oscillograph or ink recorder. In the final adjustments, unbalance at operating speeds is also observed. The artificial line adjustments have their limitations (which are not usually the same for different cables), because of inadequacy of artificial line, interferences, or other factors.

Corrective resonant networks are designed to secure the ultimate duplex balance needed for satisfactory operation. These networks, Fig. 4, supplement the artificial line and provide an impedance, which can be effectively controlled as to magnitude and width of frequency band. The networks are designed in accordance with a frequency characteristic of the residual unbalance.

This frequency characteristic is obtained by taking electrical measurements, directly across the duplex bridge over the important frequency range, of the effective residual impedance unbalance at the head of the artificial line. Several of the single networks are usually used, some of them being inserted in series with the cable and some of them in series with the artificial line and adjusted, as required. Final adjustment is made in the usual artificial line controls.

Figure 7 shows the frequency characteristic of unbalance on a transatlantic cable before and after insertion of the corrective resonant networks and the final adjustments. The balance has been improved by these networks over the most important frequency band, with the greatest improvement at the higher frequencies.

Crossfire between cables usually occurs if two or more cables land at the same point and extend underground to the cable station with small separation. This interference is controlled by applying simple corrective networks at the cable station. If two cables land at different points and are laid closely parallel for some distance, the corrective problem is more difficult, requiring more complex networks.

With the techniques that have been developed for improved duplex balances and the correction of crossfire, extraneous interference, mostly of natural origin, becomes controlling in signal speed. Natural interference is picked up in the shallow-water end sections of cables, its magnitude varying widely from cable to cable, depending on the depths and distances encountered. Receiving earths of the non-loaded cables are usually located up to several miles from shore. The twin-conductor end sections of cable greatly assist in limiting the interference and could be extended further into the ocean, but the cost would be correspondingly increased.

Some improvement in interference levels has been obtained by increasing the sending voltages, restricted until recent years to about 50 volts, up to 90 to 120 volts, now in common use by some of the cable companies.

Signal-shaping amplifiers of the type just described are being used on many of the longer transoceanic cables; short-cable amplifiers are employed for short connecting cables.

On certain duplex-operated non-loaded cables, the sum of the letters-per-minute speeds in the two directions now averages in printer operation over 40 per cent more than with the previous recorder speeds. The average net increase in the message capacity of these cables, considering short-cut methods permissible in recorder but not printer operation, is over 30 per cent. In one case, it is expected that the ultimate speed in printer operation when three-channel multiplex equipment is available will be 750 letters per minute in each direction of transmission.

In loaded cables, the channel speed is usually 250 to 300 letters per minute, the number of channels per cable varying between 4 and 8, depending on the cable make-up. One loaded cable, about 2300 nautical miles long, operates at 65 cycles per second and provides 5 one-way channels, each with a capacity of 312 letters per minute.

Power equipment consists of storage batteries and charging-equipment installations of the capacities and sizes necessary to provide both regular and emergency power at the cable stations. The d-c voltages range from 90 to 120 volts.

TELEGRAPH EQUIPMENT

By John D. Taylor

The telegraph central office is the centralizing point which, for its particular area, directs and regulates the movement of telegraph messages. From such points, the operating personnel also maintains constant supervision over the proper functioning of and needed repairs to equipment and outside plant.

13. CENTRAL-OFFICE EQUIPMENT

Central-office equipment for telegraph operation consists of a wide variety of equipment units and associated facilities. The types and amounts of equipment at any given central office vary over a wide range, depending mainly upon its importance in the general telegraph network, the message volume handled, and the nature of the traffic. Some of this equipment has previously been described and will not be discussed further here.

In a large office, the principal classifications of equipment may be considered to include:

1. *Terminal equipment for open-wire lines*—entrance cables, protector and distributing frames, and testboard (Western Union designation is line-terminal switchboard).

2. *Intermediate operating equipment*—repeaters, concentrators, and system apparatus, such as multiplex, varioplex, reperforator-switching, carrier, and Telefax.

3. *Operating positions*—teletypewriter switchboard (Bell System designation), operating tables or positions for Morse, multiplex, and teleprinter sending and receiving apparatus, tape perforators, reperforators and transmitters, Telefax terminals, telephone lines, and switching facilities.

4. *Message-handling equipment*—belt conveyors between incoming and outgoing operating positions and distributing centers within an office, and pneumatic-tube terminals from branch or other message-handling centers to the main office.

5. *Power equipment*—power-generating equipment, such as rectifiers and motor-generators, power switchboards, power distributing systems to power-operated units, storage batteries, and emergency power plants.

6. Building equipment—lighting, heating, ventilation, elevator service, personnel quarters, and many other items of this nature.

The equipment layout in an office should provide a minimum travel time of the operating personnel in their regular work and should limit wiring and cabling requirements between equipment units.

Protector (main) frames provide for termination of the entrance cables through which the open-wire lines extend into the central office, either directly on protectors or on terminal blocks, from which the lines are connected to protectors. The protectors, consisting of heat coils (or fuses) and carbon block discharge gaps, function to prevent excessive foreign currents or voltages from damaging the central-office cables and equipment, as discussed in more detail in Sections 10 and 17. Office circuit fuses are also provided to prevent excessive office currents from damaging the equipment or wiring.

Testboards or line-terminal switchboards, to which the lines are extended from the protector frames in cable, are designed to terminate the telegraph circuits, entering an office, in jacks for testing, patching, and other purposes, as required in maintaining and operating these circuits. Certain central-office equipment units, battery taps, and special-purpose apparatus are also terminated at such boards and may be associated with or disconnected from the various circuits, to meet operating needs. In the *older-type boards* telegraph-circuit layouts are usually established in part by means of patching cords; in the *latest-type boards*, these circuits are wired through groups of jacks, individual to each circuit, eliminating the need for patching cords, except for testing or establishing other than the normal circuit layout.

Intermediate distributing frames have mounted, on their two sides, terminal blocks, to which the various equipment units in the office are wired. Office cables also extend from these frames to testboards and line-terminal switchboards and to telegraph operating positions and teletypewriter switchboards, so that by means of crossconnections on these frames circuits may be connected to the various testboard and switchboard jacks, operating positions, and equipment units, as desired.

Teletypewriter switchboards employed for the purpose of establishing connections between teletypewriter subscribers consist principally of positions equipped with jacks, cord circuits, and a keyboard sending and receiving teletypewriter, which may be associated by means of keys with any cord circuit on the position.

The operator handles connections somewhat like an operator at a manual telephone switchboard position, the principal difference being that the incoming calls are answered and extended to the called subscriber by operating the teletypewriter, with the assistance of similarly equipped distant operators, if necessary.

These boards are designed to serve as few as 10 lines (mostly for private networks) or as many as 2040 subscriber lines and 600 intertoll trunks, when the outward, inward, and through traffic is handled at one board.

In order to improve transmission from a central office to a teletypewriter station, a wave-shaping network, consisting of resistance, inductance, and capacitance of various values and combinations, depending on the type of loop and connected equipment, is frequently inserted in the side of the loop connected to the repeater at the central office. Wave-shaping networks are also employed, as required, at the stations. These networks assist in restoring the received signal wave to its original shape.

The jacks, cords, and plugs used at telegraph switchboards and testboards for testing, patching, and establishing connections may be of standard types, such as those in manual telephone testboards and switchboards. However, where low-resistance conductors with greater service margins are needed, these units are frequently of heavier construction. The number of conductors will vary between different boards, depending on the circuit requirements.

Telegraph repeaters employing polar transmission are standard for d-c trunk-line terminal sets; they usually operate on a full duplex basis with ground return. Other types of repeaters are used at intermediate trunk-line points for transmission reasons. The operating functions of repeaters have been discussed in article 6.

Repeaters are used in various arrangements in circuits, the name by which they are designated indicating the manner in which they function, such as a combination duplex-duplex half repeater, terminal duplex-duplex half repeater, and high-speed polar duplex, high-speed single-line, and regenerative repeater. One form of repeater provides for receiving, recording, and, when the outgoing circuit is clear, retransmitting telegraph signals, which, in effect, is equivalent to storing signals. Special types of repeaters serve other purposes, such as connecting multiplex channels to other channels and loops.

Relays perform vital functions in the operation of telegraph circuits and equipment. The modern high-speed polar relay operates efficiently and with precision. Figure 1 shows a common type of polar relay for use in high-speed d-c telegraph circuits. Figure 2 shows a plug-type polar relay, commonly used in d-c telegraph line circuits. Many other types of relays are employed in telegraph circuits, each designed to perform its particular function.

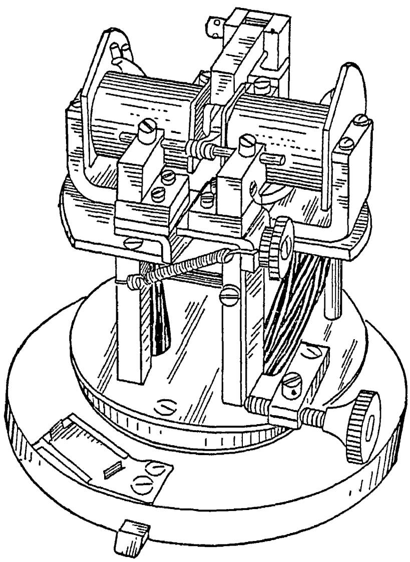


FIG. 1. Typical High-speed D-c Telegraph Polar Relay (Courtesy Western Union Telegraph Co., A.I.E.E., F. E. D'Humy, and P. J. Howe)

Multiplex distributors, as previously explained in this section, are devices having segmented and solid ring face plates, with which rotating brushes are in contact, and by means of which telegraph signals from one or more circuits are transmitted over a single telegraph line in sequence on a time-sharing basis. The face plates are removable and may be changed as desired to meet operating requirements.

The brushes are mounted on a shaft, driven by an impulse motor, which is synchronized with the motor of a multiplex set at the distant line terminal. For this reason, the start and stop pulses are not required, as they are for the teleprinter or teletypewriter, and only a five-unit code is employed per character for each circuit operating over the multiplex line.

The shaft between the motor and the brush assembly consists of two parts joined together by a magnetically operated ratchet device, by means of which the angular position of the brushes with respect to the motor rotor may be changed in steps of 1 $^1/_2$ angular degrees. The change may be made with the motor operating, if desired. A mercury-filled flywheel mounted on the motor shaft provides stability of rotation. The various connections to the distributor are brought out to multicontact bayonet-type plugs to provide for rapid replacement of the distributor in case of trouble. Figure 3 shows one type of multiplex distributor.

The *shaft speed* is determined by the required channel speed. For four-channel operation in each direction, at a channel speed of 66 words per minute, the shaft speed is about 396 rpm. However, the total message capacity for the line is 528 words per minute.

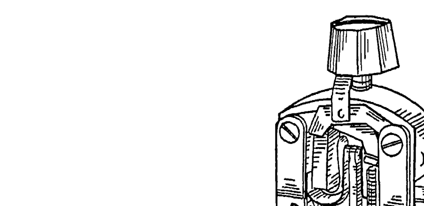


FIG. 2. Typical D-c Telegraph Relay for Line and Other Circuits (Courtesy Western Union Telegraph Co., A.I.E.E., F. E. D'Humy, and P. J. Howe)

Synchronization of speed between two multiplex sets at opposite ends of a multiplex circuit is accomplished by means of a *driving fork* associated with each set. This fork is magnetically vibrated and equipped with contacts to generate impulses from a d-c supply for operating the motor of the set. The frequency of vibration of the fork, and hence the motor speed, may be altered by changing the position of weights clamped to the fork tines. The normal fork frequency (without weights) is about 60 cycles per second, corresponding to a distributor speed of about 360 rpm. Forks with shorter tines are used for high-speed circuits. Figure 4 shows a drawing representative of a driving fork.

The **tape perforator** consists of a perforating mechanism, actuated by a keyboard unit, in which each individual key lever, with certain exceptions, is designated with an upper-and lower-case character. By operating the key levers when the perforator is in its operating condition, the characters corresponding to the keys operated will be punched in a paper tape in the standard five-unit code.

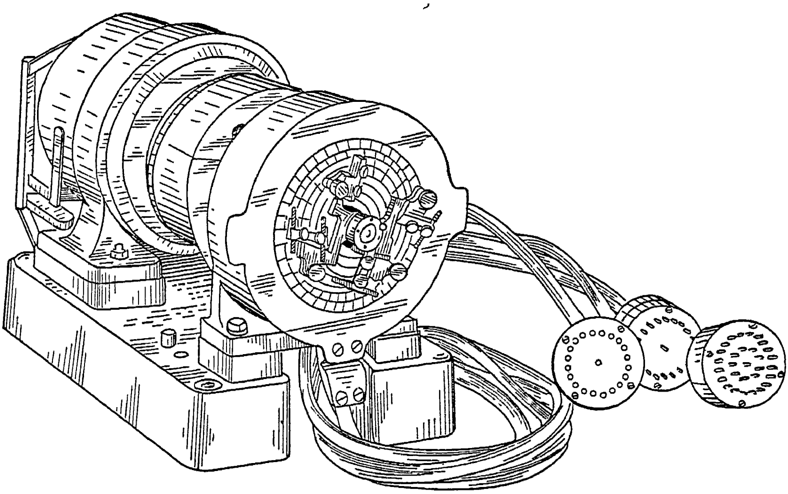


FIG. 3. Typical Rotary Distributor (Courtesy Western Union Telegraph Co., A.I.E.E., F. E. D'Humy, and P. J. Howe)

In one type of perforator design, a punch block contains six small cylindrical metallic fingers or punches, between the die plates of which block the tape is fed (see Fig. 5). A punch hammer, operated by a magnet, forces the punches through the tape as it passes the punch holes in the die plates. As each character is punched, the tape is moved forward one space by a pawl and feed roll, and the succeeding character is then punched. Five of the punches are for code perforations, and the sixth punch provides the feed holes in the center of the tape.

Between the punch hammer and the five punches are five punch bars, which are connected by bell cranks to five U-shaped bars (loops), pivoted at each end and held by

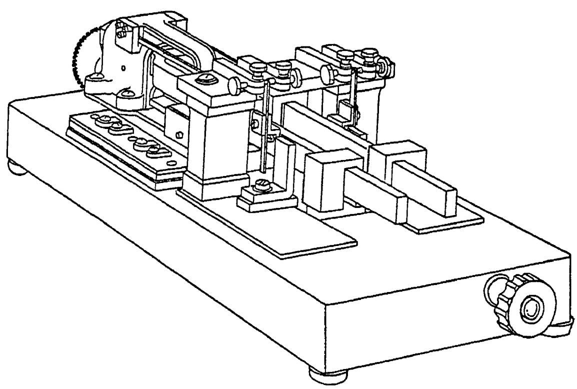


FIG. 4. Driving Fork Used in Multiplex Systems

means of springs so that their greatest length is in a horizontal position directly beneath the keyboard.

Attached to the lower edge of each key lever is a piece of metal, called a comb, which is cut out, so that depressing a key will cause its comb to strike the top edge of one or more of the loops and move them downward. The combs are cut differently for the different keys, resulting in a different combination for each key depressed.

The depression of any loop moves the corresponding punch bar from in front of its punch so that, when the punch hammer is operated by the magnet, the corresponding punch does not operate and the tape is not perforated by that punch. A sixth (power)

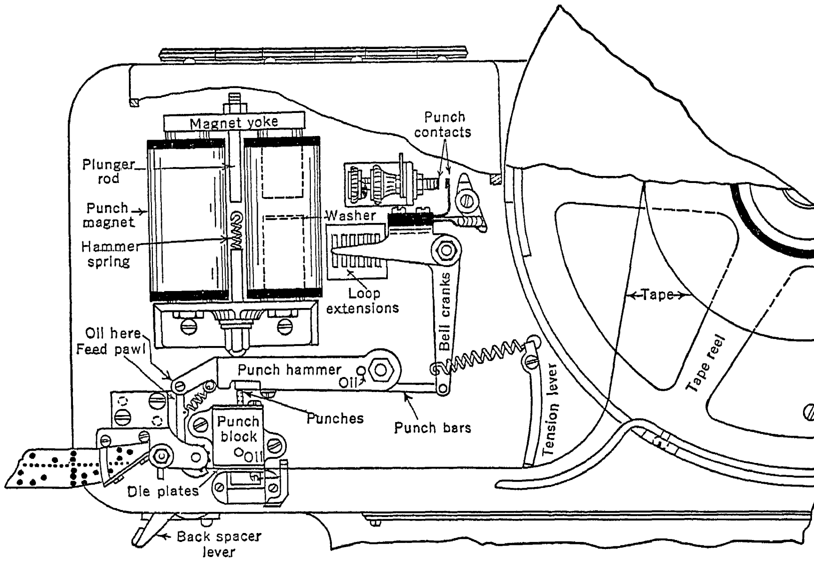


FIG. 5. Perforator Punching Mechanism (Courtesy Bell System)

loop, operated when any key is depressed, energizes the punch magnet, which actuates the punch hammer.

This perforator also provides for moving the tape backward for correction of errors and for indicating end of a line, so that the carriage return key can be operated to start a new line at the distant receiving machine.

The **typing reperforator** is a device for receiving messages from a telegraph circuit or transmitter and recording them in tape by five-unit code perforations and by printing the character on the tape above the corresponding perforations.

The **transmitter and transmitter-distributor** are devices for translating code perforations in tape into electrical impulses, which are transmitted over a connecting medium to a receiving device for interpretation as signals of intelligence. The perforations may be in a five-unit, six-unit, or other code, depending on the particular circuit requirements.

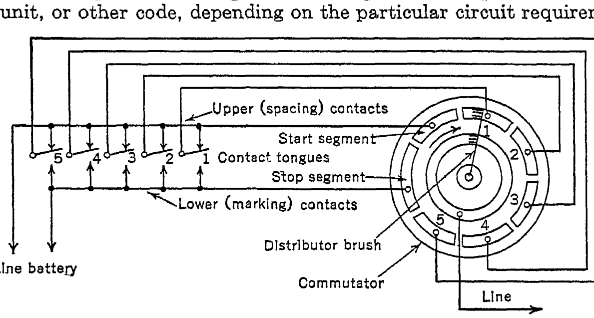


FIG. 6. Diagram Showing Transmitter Contacts Wired to Distributor Segments (Courtesy Bell System)

In the transmitter-distributor the tape transmitter establishes the code combinations to be transmitted, and the commutator distributor sends out these combinations over the line, as marking and spacing impulses, in their proper sequence and at the desired speed. Both units are driven by the same speed-regulated motor.

For one type of five-unit code transmitter-distributor, the five contact tongues (see diagram in Fig. 6) of the transmitter move between two sets of contacts, one set marking and the other set spacing. These tongues and the multipled marking and spacing contacts are connected to distributor segments. In "make-break" operation, battery is connected to the marking contacts only.

The tongues are mechanically connected to the ends of five pivoted contact levers, each of which has three extensions *A*, *B*, and *C*, as shown in Fig. 7. In the unoperated position of the contact lever bail (the position shown in Fig. 7), the contact lever springs pull down on the *A* extensions, causing the tape pins in the *C* extensions to press up against the tape but the upper contacts remain closed. Since the tape pins are spaced the same distance apart as the tape perforations, any pin will then pass through the tape if there is a perforation in the tape above it. When a pin moves through a perforation, the *A* extension is permitted to move down slightly under action of its spring, thus opening its spacing and closing its marking contact by the movement of its contact tongue. Where there is no perforation in the tape above a pin, the pin is held in its normal position against the

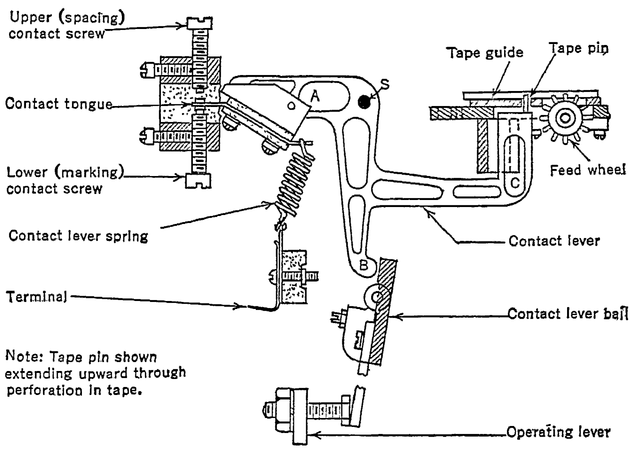


FIG. 7. Tape Transmitter Mechanism (Courtesy Bell System)

tape and its contact tongue remains on spacing. Thus, the code perforations in the tape determine the setting of each of the contact tongues either to marking or spacing, and hence the polarity of the distributor segments connected to the tongues.

After each character is transmitted, the contact tongues are reset to spacing by operation of the operating lever and contact lever bail. This bail moves the *B* extensions to the left, withdrawing the tape pins to a position below the tape guide surface, thus moving the contact tongues upward. Also, after each character is transmitted, a sixth (feed) lever is actuated, causing the tape feed mechanism to move the tape forward a distance equal to that between the character punches in the tape.

An automatic stop is mounted on the transmitter-distributor base to stop the transmitter if the associated perforator operation is interrupted or if its speed becomes less than that of the transmitter. This avoids tape mutilation by the transmitter. The stop consists of a light metal lever, suspended over the tape loop between perforator and transmitter, which is raised when the loop becomes tight, opening the control circuit of the transmitter.

Message conveyors are employed in the larger telegraph central offices to reduce the travel and handling time for messages that must be transported from one location to another in the same office or building. In the largest centers, the total number of messages handled daily may average 300,000 or more.

Pneumatic tubes provide a rapid and efficient means of transporting the original copies of messages, being commonly employed between branch and central offices and for intradepartmental use in large central-office buildings.

14. STATION EQUIPMENT

Station equipment intended for telegraph purposes is of various types and designs to meet the needs of the customer. A number of the equipment units are also applicable for use at customer premises, such as the teletypewriter, teleprinter, perforator, typing reperforator, transmitter-distributor, Telefax, ticker, clocks, and various other units. The station equipments discussed in the following paragraphs are additional to those previously discussed.

Printers. Two types of teletypewriters of the start-stop, five-unit code type have not been previously described. One is a motor-driven, single-magnet, fixed paper carriage, typebar, page printing type, operating normally at speeds of 240 to 368 operations (40 to 60 words) per minute. The paper may be in single sheet rolls, usually 8 or 8 1/2 in. wide, or two or more carbon copies may be made by using the proper paper assemblies.

The other is a teletypewriter, used where not more than one carbon copy is required and where a smaller machine than the first is desired by the customer. This machine is a motor-driven, single-magnet, moving paper carriage, typewheel, page printing type, operating normally at 368 operations per minute. It uses paper 8 1/2 in. wide, which may be multiple wound for one carbon copy.

Radio-interference-suppression apparatus (filters), consisting of inductance, capacitance, and resistance, are employed in station equipment, usually in a parallel-series relation, across various make-break contacts in the teletypewriter or teleprinter. One type of suppressor reduces induction at broadcasting frequencies, and another type provides suppression at both broadcasting and higher frequencies.

Selectors are used on important circuits to provide a convenient means for calling attendants at repeater or terminal stations to the circuit when trouble develops. They are also used on Morse wires and concentrators to enable one station to call another without calling in all the other stations on the same circuit. The selector, Fig. 8, contains a

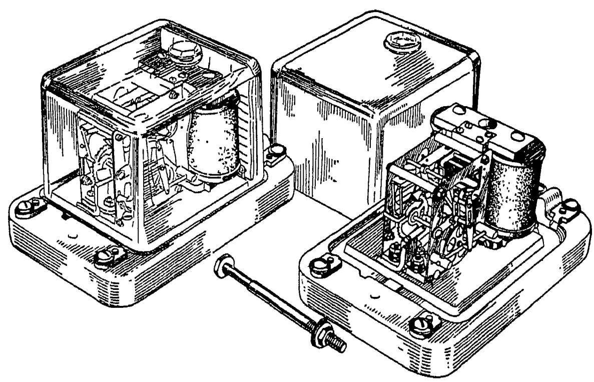


Fig. 8. Typical Telegraph Selector

magnet, normally connected with the line or line relay, which controls a mechanism arranged to close a set of local contacts only when the magnet is operated by the particular combination of impulses for which the selector mechanism is adjusted. The local contacts of the selector may be used to operate either visual or audible signals or to place the local receiving apparatus in an operative condition. Signal combinations for operating the selectors may be transmitted manually with a Morse key, or special clockwork-driven calling keys may be used for the purpose.

In Morse telegraphy manual operation is employed, but, owing to the growth of automatic transmission of messages, Morse operation is confined mainly to occasional local services over short-haul telegraph facilities, such as those to sporting-event locations, railroad-station offices, and small towns.

Morse circuits are operated either single or duplex, as conditions may require. Where several such circuits terminate at a central point, they are usually connected into a concentrating unit by means of which one operator can handle all the circuits so concentrated.

Figure 9 shows sketches of a Morse sending key, sounder, and relay.

The sending key has two contacts, normally held open by a spring, which are closed by manually depressing a key lever, which also depresses the spring. When the key is released, the spring causes the contacts to open. A second lever operates horizontally to close the circuit at the key when the key is not in use.

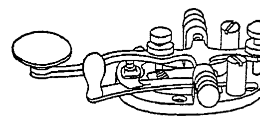
The sounder consists of an electromagnet and an armature which moves a sounding lever between two adjusting stop screws, the assembly being mounted on a sound-amplifying base. When the magnet is energized, the armature is drawn down until its stop screw contacts the metal frame, producing an audible click. When the magnet is de-energized, the armature is restored to its unoperated position by spring action, its outer end striking the upper stop screw, which produces a click.

Two of these clicks separated by a short interval are interpreted by the operator as a dot. For a longer interval between clicks (usually three times the interval between dots), the signal is interpreted as a dash. The armature travel and restoring spring tension can be adjusted to suit the operator.

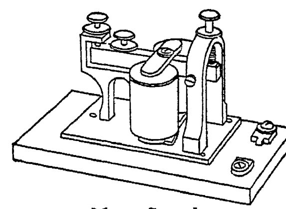
Local sounders operate in local sounder circuits; they may be of low resistance (about 4 ohms), requiring an operating current of about 250 ma, or of high resistance (100 or 400 ohms), requiring operating currents of about 60 and 30 ma, respectively. *Main line sounders* are designed for operation in series with the main line circuit. These sounders may be adjusted so that their operation is not materially affected by line leakage, usually encountered, and their resistance may be 30, 100, or 120 ohms, depending on circuit requirements.

The relay consists of an electromagnet having one or more windings arranged to move an armature, which operates between a set of contacts. Signaling impulses, passing through the operating winding of the relay, cause the armature to move to its proper contact, in accordance with the type of impulse received, thus repeating the sent signals in a local circuit connected to the relay contacts.

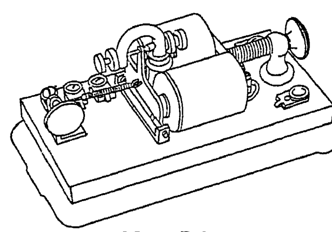
The type of relay shown in Fig. 9 usually is equipped with a magnet having 25, 100, or 150 ohms resistance.



Morse Sending Key



Morse Sounder



Morse Relay

FIG. 9. Telegraph Apparatus Used at Manual Stations

TRANSMISSION-MAINTENANCE

By John D. Taylor

15. TRANSMISSION STANDARDS

Signal transmission is considered perfect in any telegraph circuit or connection if the received effective marks and spaces or dots and dashes are exactly the same length as the sent marks and spaces or dots and dashes. In practice, signals which may be nearly perfect as sent are affected during transmission by circuit constants, such as the inductance, capacitance, resistance, and leakage of the line conductors, by equipment characteristics, and by various forms of interference. It has been found that, with present operating arrangements, bias changes, large fortuitous distortion (usually termed "hits"), and smaller but more frequent fortuitous distortion (see discussion in Article 5) are the principal causes of transmission impairment.

In designing telegraph facilities, both line and equipment, the general problem is to provide a satisfactory grade of service in the most economical and convenient manner that will meet public and operating needs. The trend is toward automatic transmission in the telegraph just as it is in the telephone field, in order to increase the speed of service most efficiently.

Because of the many types of facilities employed in the telegraph plant, different telegraph circuits affect telegraph signals differently, and, in order for the signal transmission to be satisfactory in any given circuit layout, it is necessary to know in advance what these effects will be.

Since, in the present state of the art, distance is not a limiting factor in the transmission of telegraph signals, direct telegraph circuits may be provided from any point to any other point in the world by choosing the proper equipment and transmitting media for such circuits.

16. TRANSMISSION COEFFICIENTS

One of the large communication companies in the United States in 1926 developed a system (since improved) of transmission ratings of telegraph circuits and equipment, based on signal-distortion measurements and on experience gained from operating performance. This system is based on the fact that, since the distribution of distortions follows the normal distribution law, ratings or coefficients, chosen as proportional to the mean-squared values of the distortions, could be added directly to give the overall coefficient for any combination of circuit units for which coefficients were available.

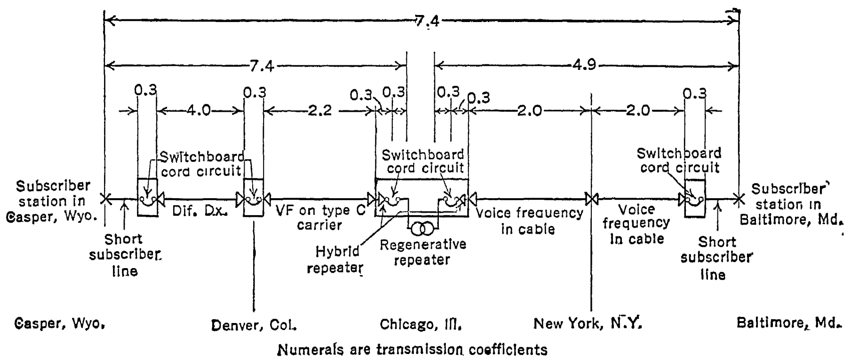


FIG. 1. Diagram of Typical Teletypewriter Exchange Service Connection Requiring a Regenerative Repeater (Courtesy Bell System)

The magnitude of these coefficients was selected so that, for satisfactory signal transmission, the overall combined coefficient should not exceed a value of 10. Where this value was exceeded in a given circuit layout, it would be necessary, by some means, to reduce the signal distortion. For start-stop telegraph equipment, the regenerative repeater is available for correcting signal distortion. It is customary to insert them in a circuit or at the junction of circuits to limit the overall coefficient to 10, as shown in Fig. 1.

Duplex telegraph repeaters of the differential type are employed at intermediate points in long d-c telegraph circuits with ground return to increase signal strength. These repeaters are usually spaced about 250 miles apart, although this distance varies over a comparatively wide range, depending on the types of line facilities, equipment, interference, operating speeds, and other factors involved. These intermediate repeaters do not correct distortion but repeat the signals through from section to section. For this reason, regenerative repeaters or other distortion-correcting devices are generally required every two or three repeater sections.

17. CROSSFIRE

Crossfire neutralization between polar duplex grounded telegraph circuits is necessary where this form of electrical interference becomes objectionable. As the speed of operation increases, crossfire between such circuits assumes greater importance.

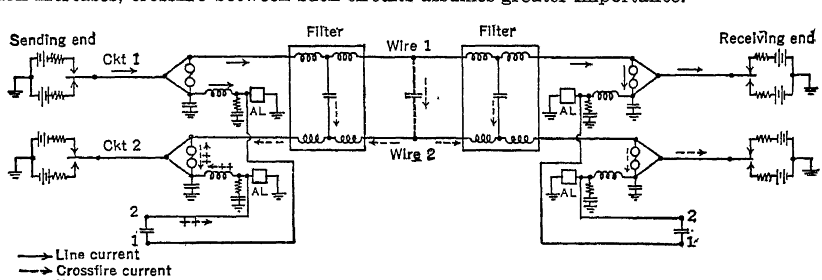


FIG. 2. Neutralization of Crossfire Current at the Sending Terminal (Courtesy Bell System)

Crossfire is caused by the mutual inductance and capacitance between the telegraph circuits and by leakage. Though the major part of this interference is due to the close physical relation between the paralleling open wires and cable conductors, some of it results from couplings in equipment common to two or more of the paralleling conductors, as in composite sets, line filters, and loading coils. In general, the magnitude of the interference is proportional to the length of the line wire and cable conductor parallel. In cable, the coupling is much greater than in open wire, and, if more than two duplex telegraph circuits are derived from one quad, the crossfire usually becomes prohibitive, even for comparatively short distances, unless neutralization is applied.

Figure 2 shows one method of neutralizing crossfire currents at the sending end without affecting these currents at the receiving end of a polar duplex circuit. Figure 3 shows one

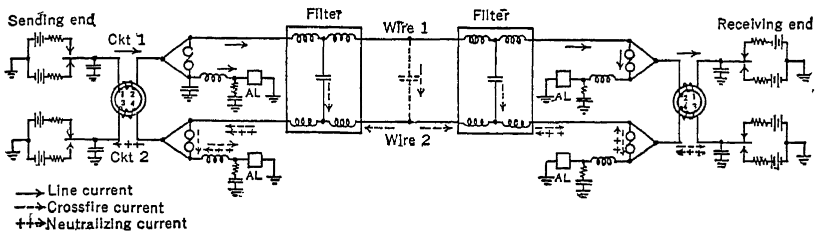


FIG. 3. Neutralization of Crossfire Current at the Receiving Terminal (Courtesy Bell System)

method of neutralizing crossfire currents at the receiving end without affecting these currents at the sending end of a polar duplex circuit. In some cases, neutralization at the sending or the receiving end only will be sufficient; in the more severe cases, both methods will be required.

For sending end neutralization, a properly adjusted condenser is employed between the artificial lines of the two circuits shown, causing currents to be set up between them in an opposing direction to the crossfire currents. For receiving end neutralization, a transformer is inserted in the apexes of the two duplex sets, which, by proper poling and adjustment of the coupling, sets up currents opposing the crossfire currents.

18. MAINTENANCE

The proper maintenance of telegraph facilities plays a vital part in providing the public with telegraph service of a satisfactory grade. For this purpose, various types of testboards and testing equipment have been developed and routines established to insure that the facilities are properly maintained.

TESTING EQUIPMENT developed for maintaining telegraph facilities includes such principal units as:

1. *Monitoring machines* with sending and receiving units, for checking the transmission of telegraph signals.
2. *Test distributors*, sending substantially perfect signals for lining up and testing telegraph circuits and apparatus. One type is for the central-office and another for portable use.
3. *Automatic multiple senders*, providing sources of battery reversals and of test signals for teletypewriter circuits.
4. *Telegraph signal biasing sets*, providing sources of biased battery reversals and teletypewriter test signals of the inverse neutral type; also providing for measuring bias from a distant sending end.
5. *Telegraph transmission stability test set*, giving quantitative indications on a recording meter of the freedom from bias of a series of received reversals.
6. *Telegraph station test set*, for testing at outlying stations or small central offices, to indicate distortion of received signals and assist in determining wave-shaping network requirements.
7. *Telegraph transmission measuring set*, indicating directly on meters the distortion in signal reversals from start-stop machines. One meter indicates bias (average) distortion and another meter the peak value of the total distortion (instantaneous sum of bias, characteristic, and fortuitous effects). One type of set is for the central office and another is for portable use.
8. *Telegraph crossfire test set*, for determining the proper values of capacitance and resistance for neutralizing sending-end crossfire and the proper inductance couplings for neutralizing receiving-end crossfire between grounded polar duplex telegraph circuits.

9. *Hit indicators*, giving locked in indications of short-duration line disturbances, sufficiently large to cause or almost cause the receiving relay to leave its marking contact, or when an initial spacing impulse of a message is received from the line.

10. *Orientation-testing indicator (portable type)*, for adjusting the orientation of a certain type of regenerative repeater.

11. *Hit suppressor unit*, for preventing certain forms of interference on private-line telegraph facilities from being transmitted in a certain direction beyond the line section or sections in which the hits occur.

12. *Carrier telegraph test set*, for maintaining carrier telegraph systems, including filament circuit tests and a drift measuring circuit for compensator relay bias adjustments.

13. *Frequency-measuring devices*, for checking carrier frequencies used in carrier telegraph systems.

BIBLIOGRAPHY

General References

1. *Bell Telephone System Practices.*
2. *A.I.E.E. Transactions.*

Special References

1. F. E. D'Humy and P. J. Howe, American Telegraphy After 100 Years, *A.I.E.E. Trans.*, Vol. 63, 1944.
2. O. E. Pierson, The Western Union Varioplex Telegraph System, *Electrical Communication*, Vol. 22, No. 2.
3. C. H. Cramer, Some Modern Techniques in Ocean Cable Telegraphy, *A.I.E.E. Technical Paper* 47–79, December 1946.

RADIO TELEGRAPH SYSTEMS

By J. L. Finch

In radio telegraph communications it is necessary to have a transmitter, and a receiver and an operator's position. For the less important circuits these are often located together under the control of one operator who alternately sends and receives messages. This method of operation is known as the simplex method. For more important circuits the transmitter and the receiver are arranged for simultaneous operation, and this is known as the duplex method. In the duplex method it is usually advisable to have the transmitter and receiver located at different points to reduce interference troubles. This arrangement becomes increasingly important as more transmitters and receivers are used. When it is necessary to locate the transmitters and receivers at the same point, as on shipboard, special precautions must be taken to prevent mutual interference and the frequency separation between transmitted and received signals must be kept relatively great.

When ratio communication is carried on to and from large cities it is the usual practice to have a central office in the city with both the transmitting and the receiving centers located well outside the city and connected by wire lines or by ultra-high-frequency radio control circuits.

19. CHOICE OF TRANSMITTER AND RECEIVER SITES

The sites are usually chosen where ample level space is available for directional antennas for each communication circuit planned. The availability of a dependable power supply and of reliable control channels is a further consideration. The receiving centers are usually located where man-made static is low. It is desirable that hills and mountains in front of the antennas do not rise at an angle of more than 3° for long-distance communication at short waves, although angles of as much as 5° can usually be tolerated. For shorter distances the angles may be greater. The land immediately in front of an antenna has a direct bearing on the vertical directivity and should be level. For long waves, mountains and valleys within a fraction of a wavelength have little effect on the antenna directivity.

20. CHOICE OF FREQUENCIES

The most desirable frequency to be used between any two points depends upon many factors which vary with time of day, season of year, position in the sunspot cycle, occurrence of magnetic storms, and location of the great-circle path over which the signals must travel with respect to the aurora areas centered about the earth's magnetic poles. These factors are mostly related to wave propagation and are described in article 10-23. Usable frequencies start at about 10 kc and extend as high as 30,000 Mc. The very low frequencies are propagated effectively and are relatively steady and consistent, but they require very large and expensive antenna systems, particularly at the transmitting end, in order to radiate even a small percentage of the generated power. At the receiving end extensive antennas are required to get good directivity and thus a favorable signal-to-noise ratio. Further, the transmitting antennas must be very sharply tuned in order to be efficient, and this fact limits their modulation capabilities to relatively low keying rates. It should be noted that propagation at frequencies from 50 kc to 30 Mc is subject to wide fluctuation.

Commercial companies have found it worth while to maintain and operate long-wave facilities which are already in existence for use during magnetic storms and to serve for re-establishing contact with the remote points after short waves have suddenly faded out.

Frequencies between 3 and 25 Mc have been found the most useful for long-distance communications. Those between 3 and 10–13 Mc are useful over these long distances when most or all of the radio path is in darkness while those from 10–13 to 25 Mc are useful when most or all of the path is in daylight. The lower of these frequencies are useful over shorter distances in the daytime. The same frequencies can be used successively in different parts of the world as the daylight and darkness areas progress around the earth.

Frequencies above 30 Mc are rarely useful for extensive periods for long distances (over 2500 miles) and those above 50 Mc can be relied upon as a rule not to carry to distant points on the earth's surface and so can be used over and over again on circuits separated by only a few hundred miles. Because signals at these higher frequencies will not bend very much, an optical path or one approaching an optical path between the transmitting and receiving antennas is necessary.

21. REDUCTION OF FADING EFFECTS

The use of short waves for communicating over long distances is very frequently beset with fading caused by the alternate addition and subtraction of the signal arriving over different ether paths. This is due to the varying phase relation of the arriving radio-frequency waves. It has been found that much of this fading is very local in character. When a signal has faded out at one location it may be coming in at full strength a few hundred feet away. Also when at a given point one radio frequency has faded out completely, another frequency only a few hundred cycles different being radiated from the same transmitting antenna may be coming in at full strength at this point. Receiving systems which employ two or more receiving antennas spaced apart geographically are known as *space diversity receiving systems*. The telegraphic signals received by each antenna are rectified and passed through a limiter which cuts off the tops of all the characters that would exceed a certain value. The various signals are then combined in such a manner that all of them must fade at once to cause a signal failure. When using three separated receiving antennas this diversity effect reduces the failures due to fading to a very small fraction of that suffered by the signal received on any one of the antennas. Systems that transmit two or more slightly different radio frequencies from one antenna are termed *frequency diversity systems*. In this case the same receiving equipment is used as otherwise, since the different radio frequencies being received are so close together that the receiver responds to all of them substantially equally. The limiting used is such that all frequencies must fade out at once to cause a signal failure. This system results in a very marked increase in the reliability of the circuit.

To obtain the different frequency components required in the transmitter output for frequency diversity it is usual to phase-modulate the carrier at an audio rate of about 600 cycles and with a maximum phase deviation of about 1 radian. This modulation results in the production of two side frequencies spaced 600 cycles each side of the carrier. The amplitude of each side frequency is a little less than half the amplitude of the unmodulated carrier. Phase modulation can be accomplished quite simply, and it does not necessitate reduction of the total power generated by the transmitter.

It is common practice to key two different transmitters operating at different frequencies with the same signals and to combine the outputs of two receivers at the traffic office to insure against service interruptions, particularly at the time of day when one of the frequencies is about to fade out and it is desired to replace it with another.

22. RADIO INTERFERENCE

Interference between radio channels should be eliminated by spacing the frequencies sufficiently far apart and by making the receivers selective enough to differentiate between the desired and undesired signals. The radiation at unauthorized frequencies such as at harmonics of the desired frequency wave must be avoided. In some types of transmitters, frequencies lower than the desired ones are generated and then the desired harmonic is radiated. The radiation of these lower frequencies must be avoided. Vacuum-tube transmitters are prone to generate spurious or parasitic frequencies. The radiation of such frequencies must be avoided. These parasitic frequencies often modulate the desired frequency wave, resulting in modulation sidebands which will be radiated with the desired wave if allowed to exist and will cause interference on other channels. Often when a transmitter is free from parasitic oscillations in the steady state, both marking and spacing, such oscillations occur during the transient period at the beginning and end of signaling characters. These oscillations are known as key-click parasitic oscillations. Other interchannel interference sometimes results from the modulation products or side frequencies generated by the keying, these being particularly noticeable at high signaling speeds and when the characters are square ended. To round the characters to reduce these side frequencies when using efficient class C amplifiers in the transmitter and with on-off keying is difficult and expensive. When using "frequency shift" or "two-tone" keying, however, the signal characters can be rounded quite simply and the objectionable side-frequency radiations greatly reduced. Two transmitters at the same station, particularly when operating at closely spaced frequencies, sometimes intermodulate each other and radiate signals in adjacent bands. This can usually be made negligible by reducing the cross coupling between the two systems.

Receivers may cause interference with other receivers, usually due to radiation from the first heterodyne oscillator into the room and into the receiver power wires and into the receiving antenna. It is particularly important to avoid radiation back into the antenna of a level of more than 1 microvolt when one antenna is used with a number of receivers at various and variable frequencies.

23. FREQUENCY SHIFT KEYING

It has been the general practice in the past to key transmitters by interrupting the transmitter output power. Recently equipment has been developed to take advantage of the "frequency shift keying" system, also known as the "two-tone keying" system. This method employs one radio frequency for "mark" and a second radio frequency for "space," each at the same power. The separation between the two is not critical. At present a separation of 850 cycles is in common use. To reduce key-click interference in adjacent channels it is desirable to shift the frequency of a single oscillator and at a rate only sufficient to accommodate the required keying speed.

In receiving frequency shift signals, amplitude variations are limited out by the use of a limiter stage. The resulting signals of fixed amplitude and variable frequency are impressed upon a discriminator the output of which may operate a relay device, tone keyer, or other utilization device. The limiter stage largely eliminates the effect of atmospheric or other interference whose strength is lower than 3 db below the signal.

Results have indicated that in practice the transmitter power can be reduced as much as 10 db, after adopting the frequency shift keying system, without degrading the service.

24. TRAFFIC OFFICE AND EQUIPMENT

In central offices carrying small volumes of traffic, particularly over radio circuits of inferior reliability, it is advantageous to have the transmitting and receiving positions located adjacent to each other. This is for convenience in asking and giving acknowledgments and for "breaking" the circuit. For handling larger volumes of traffic over a number of reliable circuits, it is advantageous to arrange the transmitting positions in one group and the receiving positions in another.

Intelligence can be transmitted over radio telegraph circuits in the Morse code or by means of automatic printers using their own special codes. For the simplest equipment Morse signals are transmitted by hand and received aurally and transcribed by hand or on a typewriter. For higher speeds and more reliable service automatic sending and receiving is used. The messages are punched on a typewriter-like machine known as a *perforator*. A sample tape is shown in Fig. 1. The tape is run through an *auto-head* device which has electrical contacts and which form the Morse characters corresponding to the letters punched. The received signals operate an ink tape *recorder*. These recorders commonly involve a moving tape with a pen pressed against it. The pen is retained in its lower position for "space" and in its upper position for "mark," making a record as shown in Fig. 1. Automatic Morse communication is practical at speeds of 20 to 500 words per minute.

For average speeds up to 40 or 50 words per minute one operator at the transmitting end perforates the tape and tends to the auto-head machine. At the receiving end one operator views this tape as the message is recorded and transcribes the messages on a typewriter. For higher speeds of transmission the tape must be punched by two or more operators and sections fed successively into the auto-head. The recorder tape must be divided between two or more receiving operators for transcription.

Automatic printers are advantageous in that they make it unnecessary for the operators to learn the code, they reduce human errors in transcribing from code signals, and they permit automatic transcription of incoming messages directly in a form suitable for delivery to the customer. Their use, however, eliminates the possibility of detecting errors

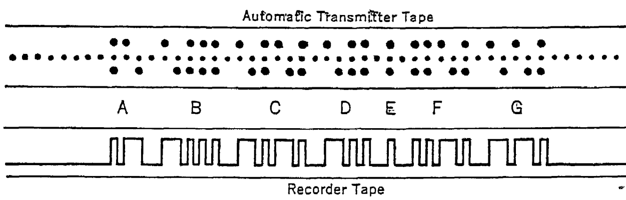


FIG. 1. Tapes for Automatic Sending and Receiving

due to mutilation of the signals introduced in the radio path. Frequently a trained operator can transcribe correctly a mutilated signal such as would cause a printer to make an error or to operate an error-indicating device. Further, the transmission speed is limited to that at which a printer will function properly, i.e., to speeds of 60 to 100 words per minute. The printers in general use employ a five-unit code and respond to any combination of signals whether mutilated or not. An error-indicating printer is in use which employs a seven-unit code. All operations use three marking and four spacing elements. Any other combination will operate an error-indicating device.

In connection with radio telegraphy it is customary to rate transmission speeds in words per minute, each word consisting of an average of five letters. More accurately, each word consists of 48 units * or minimum elements, each unit being taken as the length of one dot or one space between dots, and the length of dashes and spaces between letters being equal to three units. From this it can be calculated that when the auto-head is running at a rate to make 40 dots per second the transmission rate will be 100 words per minute. Similarly printer speeds are calculated on the basis of six operations per word, allowing five characters and one space per word.

MULTIPLEX. In order to make use of the high signaling speed capabilities of short-wave radio circuits and still enjoy the advantage of automatic printer operation at nominal speeds, means have been devised for dividing the total time available for transmission between two or more channels. This system is known as time division multiplex (see article 7). This system can also be used for Morse operation; it is advantageous because each channel can be copied as it comes in, and thus delays in message delivery caused by the necessity of dividing tape received at high speeds between two or more operators can be avoided.

MULTICHANNEL. A second method of obtaining more than one telegraph channel on a single radio circuit utilizes single-sideband equipment such as has been developed primarily for telephone service. Such a system can carry a large number of individual tones, each keyed as desired with printer signals or Morse signals and each separated from the others at the receiving end by means of wave filters.

Both the multiplex and the multichannel systems are more expensive initially and more costly to maintain and operate than single-channel systems, but they are worth while when large volumes of traffic are to be handled because they save space in the radio spectrum and at the same time provide for printer operation or Morse operation at speeds that can be transcribed currently.

* The term "baud" has, in some sections of the industry, been applied to these units. In the officially correct usage the baud is the unit of telegraphic speed or of rapidity of modulation corresponding to one minimum element per second.

25. CONTROL CHANNELS

The signals must be carried from the central office to the transmitting station, and from the receiving station to the central office, over suitable control channels. These channels can be carried over wire lines or over ultra-high-frequency radio circuits. When wire lines are available d-c signals may be used, either unidirectional or polarized (plus and minus), or keyed tone signals may be employed. The tone signals have the advantage that a number of them can be carried over one tone pair and separated by wave filters (see article 11-10). When ultra-high-frequency radio control circuits are used, keyed tone signals are employed. A number of similar groups of tone signals can be transmitted over a single radio circuit by a system similar to the voice-carrier system as described in article 17-9.

TONE KEYERS. The keying device or tape transmitter device in the central office usually closes an electric circuit by means of a pair of contacts for marking and opens it for spacing. This action is made to key a tone by means of a tone keyer such as that shown in Fig. 2. The equivalent results can be accomplished by purely electronic means

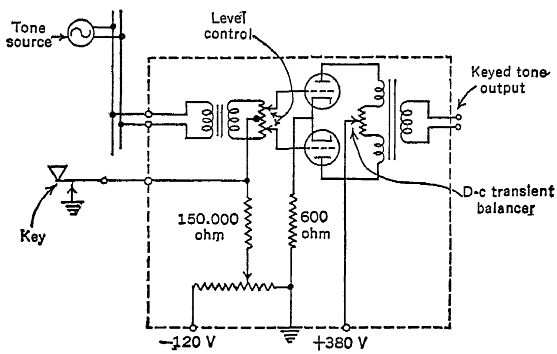


FIG. 2. Circuit Diagram of a Tone Keyer

applying the same general principles. Tone keyers are also used at the receiving station by means of which the incoming signal is made to key a tone for transmission to the central office.

TONE SIGNAL CONVERTERS. At the transmitting station the tone signals must be converted into d-c signals to key the transmitter. The device for accomplishing this is known as a *tone signal converter* and may take the form shown in Fig. 3. This embodies a

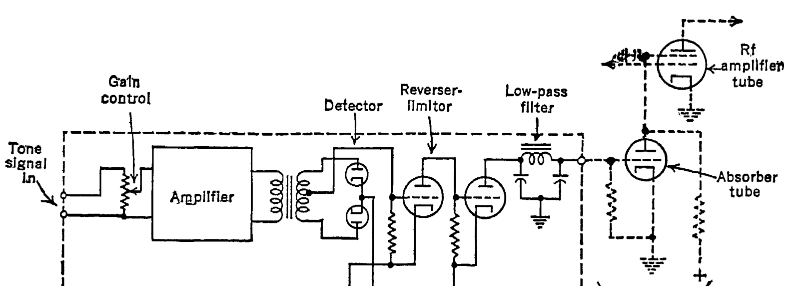


FIG. 3. Circuit Diagram of a Tone Signal Converter

threshold device for suppressing low-level noise and for preventing tails on the characters from causing "heavy" signals. It embodies a *limiter* for cutting off the tops of the signals and a low-pass filter to smooth out the rectified tone and thus to keep the signals from being modulated by the tone itself. The gain control is used for setting the level of the tone signals, which usually have been rounded by passing through the wave filters of the tone channel to give the correct "weight" of signals in the transmitter. Lowering the gain makes the signals lighter, and raising it makes them heavier.

RECORDER DRIVE. At the central office the incoming tone signals must be converted to direct current to drive the ink tape recorder or printer relay. The device for accomplishing this may take the form shown in Fig. 4. A simple tone rectifier circuit will usually suffice for driving the printer relay for printer service.

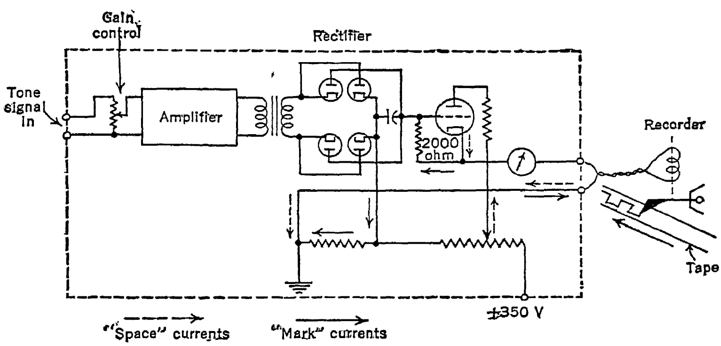


FIG. 4. Circuit Diagram of a Recorder Driving Unit

BANDWIDTHS. The transmission of keyed tone signals requires a definite band which must have a width roughly proportioned to the keying speed. Satisfactory service may be achieved by allowing for the transmission of no more than the first side frequency of the keying speed under stable input conditions and stable transmission characteristics such as are normally achieved in first-class lines. Thus for 125 words per minute Morse, i.e., a keying frequency of 50 dots per second, the usable channel width must be at least 100 cycles. For unstable conditions it is desirable to transmit the first, second, and third side frequencies. Thus, a 100-cycle band width will be suitable for speeds of 42 words per minute. In actual practice 100-cycle bands have been found suitable for speeds up to 60 words per minute.

SECTION 19

FACSIMILE TRANSMISSION AND RECEPTION

BY

MAURICE ARTZT

FACSIMILE TRANSMISSION AND RECEPTION

By Maurice Artzt

SCANNING SYSTEMS

Facsimile is defined to include all systems whereby a picture is broken into separate picture elements, these elements being transmitted by some connecting means to a distant recorder where they are reassembled into their original positions to form a copy of the original. The word "picture" in the above statement includes also diagrams, typing, handwriting, photographs, and any other form of printed or written material.

Three distinct operations are performed in the transmitting and recording of facsimiles: first, the breaking up of the picture in some orderly manner into its separate elements of varying shades, this process being called scanning; second, the transmitting of these elements to the recorder by means of signals arranged to represent the electrical equivalent of these elements; third, the rebuilding of these signals by a recorder into a printed copy of the original by a reversal of the scanning process.

A fourth part of a facsimile system, supplementary but very necessary, is a method of synchronizing the recorder and scanner. The timing of the signals received must agree exactly with the timing of the recorder, in phase as well as frequency, if the copy received is to be undistorted.

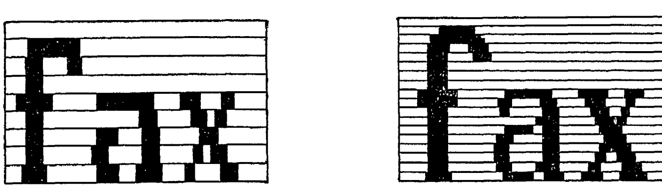
In the following articles the terms used are in accordance with the definitions and standards as set up in 1942 by the Institute of Radio Engineers. See the first reference in the Bibliography.

1. PICTURE ELEMENTS

In processing a picture by facsimile, the picture is resolved into dots, or picture elements, similar to the small dots used in printing a picture in a newspaper or magazine. These dots are obtained by "screening" in the printing process; they are obtained by scanning in facsimile.

Halftones in newspaper work have from 60 to 120 dots, or picture elements, per inch, whereas fine magazine printing may use as many as 250 dots per inch. In facsimile the limits are of about the same order, almost all present facsimile systems using 100 dots (or lines) per inch, as an average. Each picture element in a facsimile is sent as a separate signal. If the number of dots is too high, the speed of transmission is very slow; if too few elements are used, the detail will not be good enough. To send a picture of 100 dots per inch requires as many as 10,000 separate signals per square inch of surface covered.

Figure 1 illustrates the difference to be expected between 50 and 100 lines per inch, when the subject matter is ordinary typing. Some of the type would be unreadable if only 50



(*a*) 50 lines per inch (*b*) 100 lines per inch

FIG. 1. Difference in Detail for 50 and 100 Lines per Inch for Typewritten Letters

lines per inch were used, as can be seen by the poor formation of the *a*. The 100-line-per-inch detail, though not forming perfect letters, leaves no doubt as to their identity. In commercial facsimile, letters from ordinary typewriters often comprise the original, and approximately 100 lines per inch are necessary to insure readability.

Dots per inch and lines per inch are used interchangeably in the above paragraph, a practice which is not always permissible. Some facsimile systems break the subject up into dots and send a separate signal for each dot, whether white, black, or gray. Others, however, break the sheet up into parallel lines and send signals only for the black areas as encountered. Each line is then a continuous signal, varying in intensity with the shading of the original, and not made up of an exact number of picture elements as the dotted picture is. The detail limits are the same in either case, and the maximum number of picture elements per square inch is the same.

These picture elements, as observed by the scanner in the process of transmission, will be of two general types, either of the simple black-and-white variety, such as typing, line drawings, and so forth, or of the halftone variety, in which all shades of gray from white to black may occur. Two separate types of scanners are not necessary, but the amplifier equipment will sometimes be different. Any system capable of transmitting and recording halftones will also operate properly on a purely black-and-white original, but the reverse is not necessarily true.

2. SCANNERS

A facsimile scanning system includes an optical-mechanical scanner designed to project a small spot of light on the subject copy, to gather the reflected or transmitted light from

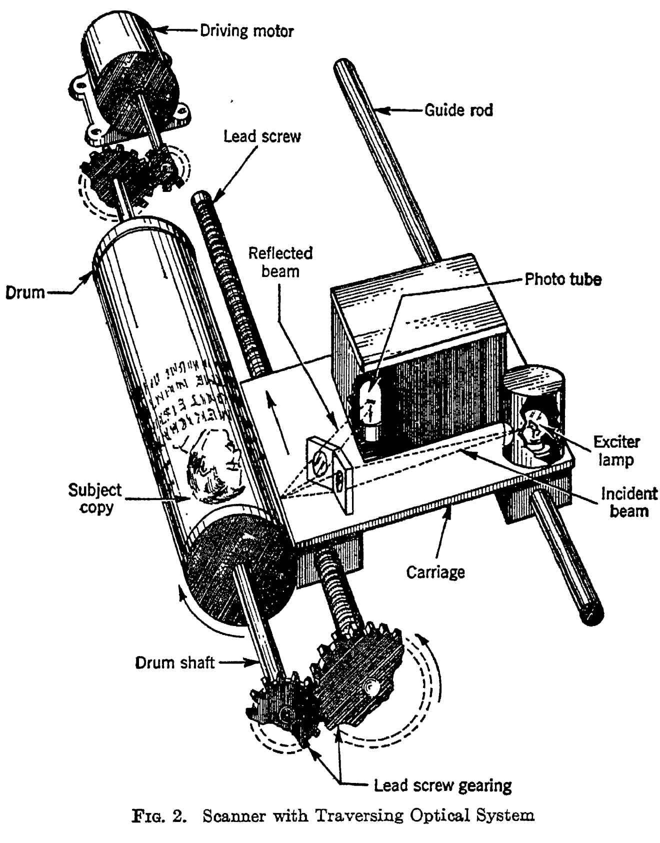


Fig. 2. Scanner with Traversing Optical System

the subject into a phototube, and to bring all parts of the subject under this scanning spot in some orderly manner. The signals generated in the phototube by the varying light values reflected from the copy are then either amplified directly or processed in other ways to form a usable electrical signal of a type suited for the particular application.

SCANNING METHODS. In the simplest form of scanning, regular lines are "ruled" across the sheet by this spot of light at some particular number of lines per inch, and signals are sent out representing each small area as it is encountered. The sheet is thus broken into a number of narrow lines, all of the same width, and these lines are transmitted one after another until the entire subject has been covered.

Scanning is generally done in only one direction and seldom back and forth. There are two reasons for this: first a unidirectional scanner is simpler to construct and requires less precision in gearing, and second the synchronizing system for back-and-forth scanning must be far more accurate.

The simplest form of the scanner, and therefore the one most generally used, consists of a drum upon which the original subject matter is wrapped, and an optical system arranged to project a small spot of light on the surface of the paper. This spot is usually somewhat smaller than the width of one scanning line. As the drum is revolved, the optical system is moved relative to the drum the width of one scanning line for each revolution of the drum. The entire subject is thus gradually passed under the scanning spot. See Fig. 2.

A phototube is arranged to pick up the light reflected from the surface of the paper, and this light reaching the phototube will be varied in intensity by the different areas of black, gray, and white that may be presented to view. The output of the phototube will be a minimum for black and maximum for white and will represent electrically the scanning of the copy. This phototube output is then applied to the input of the amplifier system.

All motions in the scanning process pictured in Fig. 2 are relative. Thus the optical system may be rotated in place of the drum, and the motion along the axis may be made by moving either the drum or the optical system relative to each other. All methods of bringing about this relative motion have been used.

In one commercial type of scanner used for news picture transmission the drum is revolved and moved along the axis by a lead screw cut on its shaft, and the optical system is stationary. See Fig. 3.

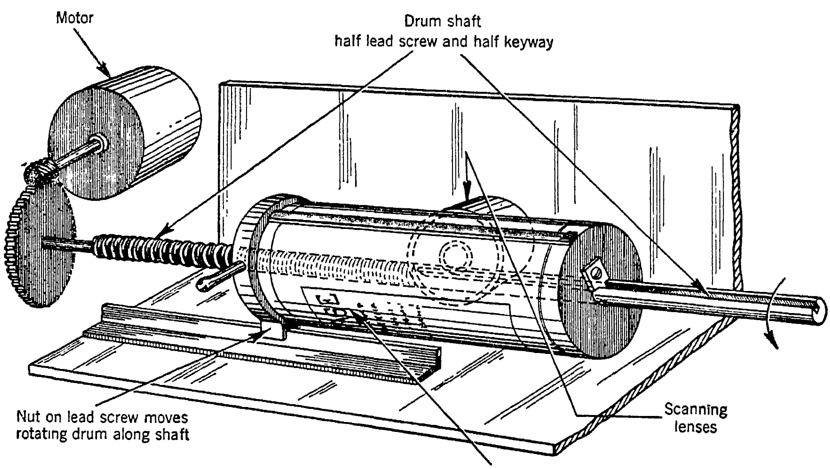
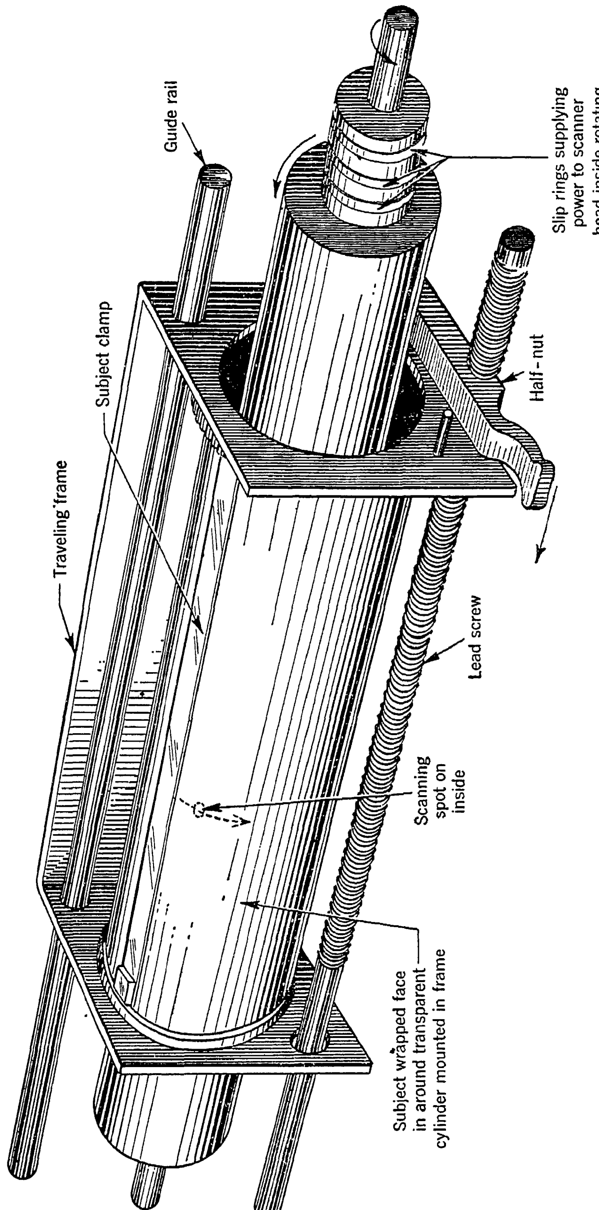


FIG. 3. Scanner with Stationary Optics and Drum Feeding for Line Advance

As facsimile speeds have increased, the time required to load the subject on a drum has become an increasingly greater proportion of the total time of transmission of the copy. Various ingenious methods of loading and scanning have been devised to minimize this loss of time. In one form of scanner designed for rapid loading, the subject is wrapped face in around a transparent cylinder, and the optical system is rotated inside this cylinder. See Fig. 4. By this method the scanning process does not have to be stopped to remove one subject and put another in place. Thus the time to resynchronize and rephase for the next subject is not lost, and the time between succeeding subjects is reduced.

In all scanners illustrated thus far the original must be of such size that it can be properly clamped on the scanning drum. Thus copy width must be approximately the circumference of the drum, minus the separation between clamps. The length is not so restricted and may be anything up to the length of the drum. Another type of scanner in which width is not restricted except as to a maximum value is shown in Fig. 5. The subject is placed face in on a stationary transparent semicylinder and held there by a curtain (not shown). Two microscopes and light pick-up systems are rotated and traversed inside to scan the copy. The two optical systems are set exactly 180° apart and in the same plane so that the signals



4

Fig. 4. Scanner with Internal Optical System and Drum Feeding for Line Advance

generated in the single phototube will be the equivalent of that from a single optical system scanning a complete cylinder. With this scanner, unloading and reloading takes only a few seconds as the copy is not clamped to hold it in place. As two optical systems are used the shaft speed will be one-half that of an equivalent drum scanner.

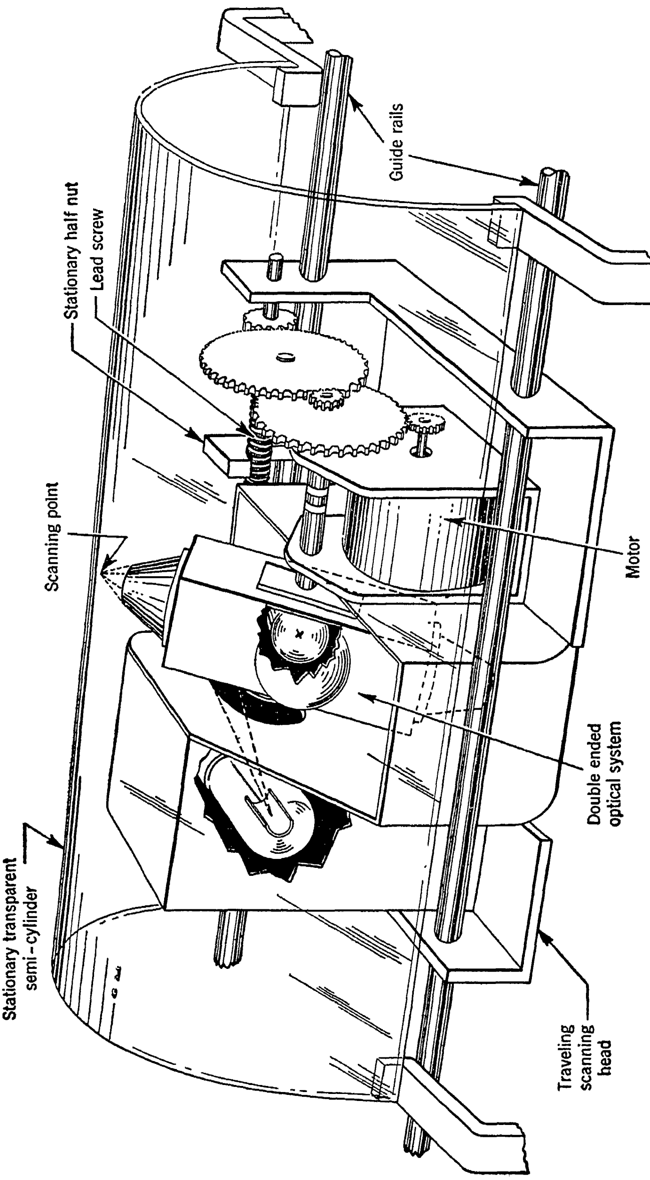


FIG. 5. Scanner with Internal Optical System Feeding for Line Advance

In one other type of scanner used for telegraph service reloading with the next message is accomplished by dropping out the drum and copy when scanning is completed and reloading with another drum containing the next message from a hopper feed. This is done automatically, and provision is made for accommodating a number of additional message-carrying drums so that the messages follow one another in rapid succession.

3. SCANNER AMPLIFIERS

The signals generated by the phototube will vary in amplitude with the shading on the subject being scanned, and in frequency with the speed of the scanning spot and the kind of subject copy. The highest frequency will be determined by the size of the smallest "dot" it is expected to transmit; the lowest frequency will be zero or a direct current to represent the large areas of white or black encountered in nearly all types of copy.

FREQUENCY SPECTRUM. The highest frequency is determined as follows: Take the width of the smallest "dot" it is expected to transmit and rule a pattern of lines of this width, separating them by the width of the dot. If the scanning is to be at 100 lines per inch then these lines will be 0.01 in. in width, 50 to the inch, and separated by 0.01 in. Such a pattern is shown in Fig. 6*A*. The fundamental keying frequency of the phototube current

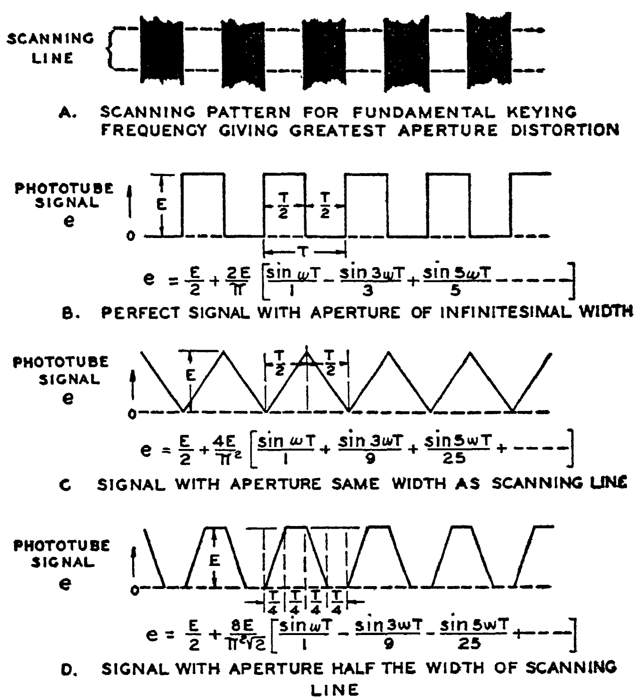


FIG. 6. Aperture Distortion of Signals

when scanning such a pattern would be 50 cycles per inch per second of spot speed. If the scanning line were 9 in. long and the drum speed 100 rpm, the fundamental keying frequency would be 50 x 9 x 100/60 = 750 cycles per second.

Higher harmonics of this fundamental keying frequency will be present as the subject scanned is a square wave pattern. If the aperture of the scanner were infinitesimal in width along the scanning line the phototube signal would be a square wave as in Fig. 6*B* and would be very rich in harmonics as shown by the Fourier series under this figure. This perfect signal is never realized in practice, nor is it desirable, for the greatly increased band width needed for transmission is not justified by the small increase in recorded detail over that obtained by carrying only the fundamental keying frequency.

As the aperture is made wider the higher harmonics become less important. When the aperture is the width of the scanning line the triangular wave in Fig. 6*C* is obtained. Here the fundamental is 81 per cent and the third harmonic 9 per cent as compared to 127 and 42.5 per cent for the square wave. When the aperture is one-half the width of the scanning line, a condition normally used in many scanners, the wave in Fig. 6*D* is obtained where the fundamental is 115 per cent and the third harmonic 12.75 per cent. When using this size of aperture very little difference can be noticed in the recorded copy whether the third harmonic is carried or suppressed. For a more complete analysis of scanning see Section 20, Television.

Almost all facsimile systems therefore carry only through the fundamental keying frequency as the upper limit of the band necessary for transmission. In the illustration given where this fundamental frequency was 750 cycles the signals from the phototube would have a frequency spectrum of 0 to 750 cycles per second. This then is the input signal to the scanner amplifier system, and, as the light reflected into the phototube is small, the input signal is usually very low in amplitude.

TYPES OF SCANNER AMPLIFIER SIGNALS. The type of amplifier used in either amplifying or processing the phototube signal will depend on the types of signal to be used in transmission, and this in turn will be governed to some extent by the transmission medium, whether it be wire line or radio. As the lower frequency limit is zero, ordinary a-c coupled audio amplifiers cannot be used. In order to carry this zero frequency or d-c component, the phototube signal is usually caused to modulate a carrier wave either in amplitude or frequency. For transmission over wire lines the carrier frequency will be chosen just high enough to carry the highest keying frequency. For radio transmission the radio carrier itself can be amplitude or frequency modulated directly by the facsimile signals, or a phone-type transmitter may be used and the facsimile modulation carried on an audio subcarrier as for wire line transmission. Finally, for short distances over wire line where no repeater stations or coupling transformers are used, a straight d-c amplifier may be used between phototube and line. As it is difficult to maintain drift-free operation of a d-c amplifier with sufficient amplification, this last method is seldom chosen.

The signals transmitted to the recorder may thus be of any one of the following types:

1. Subcarrier amplitude modulation (SCAM).
2. Subcarrier frequency modulation (SCFM).
3. Radio carrier amplitude modulation.
   a. Direct without subcarrier.
   b. With SCAM.
   c. With SCFM.
4. Radio carrier frequency modulation.
   a. Direct without subcarrier.
   b. With SCAM.
   c. With SCFM.
5. Direct-current signals.

For wire line transmission, signals of type 1 or 2 are commonly used, with 5 occasionally on short control lines, for instance between scanner and radio transmitter. It is more usual, however, if d-c signals are wanted to control a radio transmitter, to transmit signals of types 1 or 2 on the control line and detect to obtain the d-c signals for control purposes.

For radio transmission over long distances signals of types 3*c* or 4*a* are normally used as they give the most reliable results. For short radio circuits, as for instance local coverage for broadcast facsimile service, 3*c*, 4*b*, or 4*c* can be used on existing voice transmitters.

Signals of types 3*a* and 3*b* are unreliable except for very short distances and have generally been supplanted by 3*c* if an a-m radio transmitter is used.

TYPES OF SCANNER AMPLIFIERS. It can be seen by the general usage of the various signals that the two important types of amplifier systems will be either for amplitude-modulating a subcarrier or frequency-modulating a subcarrier. Where d-c signals are required it is customary to use either one of these and then detect after amplifying to the desired level.

SUBCARRIER AMPLITUDE MODULATION METHODS. There are three general methods of obtaining signals of this type. First, from the standpoint of the length of time it has been in use, is scanning with chopped light. If the light reaching the phototube is made to flicker, either by modulating the light itself or by using a mechanical shutter or chopper, the output voltage developed by the phototube will not be a direct current but a pulsating voltage which may readily be amplified by an ordinary a-c amplifier.

The minimum frequency of chopping will be determined by the fundamental keying frequency of the scanning process; the chopper frequency will be the carrier frequency and must be high enough to carry the shortest "dot." In practice the chopper frequency is usually made between 2 and 3 times the fundamental keying frequency, with $2^1/_2$ a good average. In the previous example, with a keying frequency of 750 cycles per second, a carrier or chopper frequency of about 1800 cycles would be used. The total bandwidth transmitted would therefore be 1800 ± 750, or 1050 to 2550 cycles. This bandwidth is narrow enough to be carried on regular voice telephone circuits and is so used for many news picture transmissions.

When an ordinary incandescent lamp is the light source, the chopper can be either a rotating disk with holes or slots, or a ribbon or reed vibrating in a magnetic field. When a glow-discharge lamp is the light source, such as used in some types of recorders, the light can be modulated directly without mechanical shutters. The same lamp can thus serve for scanning or for photographic recording on the same machine.

Two other methods of obtaining a-m signals are shown in Figs. 7 and 8. In the first the carrier is fed to the special phototube in a balanced bridge circuit, and a balance for minimum output signal is obtained by means of the variable resistor and capacitor with the phototube dark. As light from the subject copy increases, the tone output of the phototube increases and a true modulation results. A simple audio amplifier to build up the modulated signals to the required level follows the phototube circuit shown.

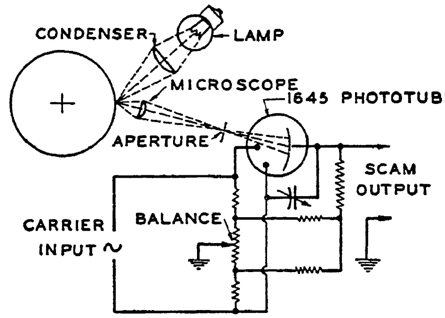


FIG. 7. Subcarrier Amplitude-modulated (SCAM) Signals Obtained by Balanced Phototube Circuit

In Fig. 8, the carrier is fed at 180° phase difference to the two screens of a pair of screen-grid tubes. The plates are connected together so the outputs of the two tubes oppose each other. By proper balancing of grid biases the output tone may be balanced to zero with the phototube dark. As light to the phototube increases, the bridge is unbalanced and the difference in output of the two tubes is obtained. With this circuit it is also possible to balance for minimum signal with maximum light for white on the phototube, and the output tone will then be a maximum for black. Thus either positive or negative modulation may be obtained by shift of the balance adjustments.

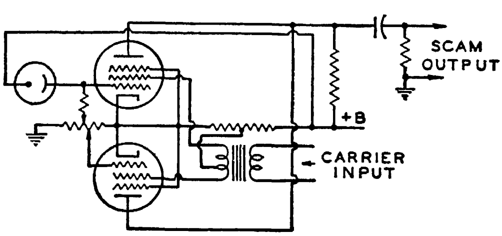


FIG. 8. Subcarrier Amplitude-modulated (SCAM) Signals Obtained by Balanced Modulator

In either of these circuits the frequency of the introduced carrier must be high enough to carry the maximum keying frequency, as explained previously with the light chopper systems.

SUBCARRIER F-M METHODS. In this type of signal black is transmitted at one frequency, white at some different frequency either higher or lower than that for black, and intervening shades of gray at proportionate frequencies between these two limits. A more complicated relationship exists between carrier, keying frequency, frequency swing, and bandwidth required than with the a-m subcarrier. This is, of course, a true frequency modulation and follows the same rules on sidebands as frequency modulation on a radio carrier. (See Section 8 on frequency modulation.) However, it has been found that, with a unity ratio of maximum keying frequency to the total frequency swing from black to white, the usable bandwidth will be confined to about the same overall limits as the a-m subcarrier with both upper and lower sidebands. Again, as with SCAM, the lowest carrier frequency must be high enough to carry the shortest dot, so the low end of the carrier swing should be at least 2 times the keying frequency. For the example with a keying frequency of 750 cycles, the carrier may swing from 1500 to 2250 cycles in going from white to black, and the total band spread (for all side frequencies greater than 10 per cent) will be from 1125 to 2625 cycles. The mid frequency for middle gray will be 1875 cycles, and only the first side frequency of $\pm$ 750 cycles need be carried.

This is only a deviation ratio of 0.5 at the highest keying frequency, but for all other picture frequencies the ratio is higher and for solid white backgrounds is practically infinite. The signal-to-noise improvement over SCAM averages at least 12db for the usual subject matter transmitted. Still greater improvement in signal-to-noise ratio would result from increased swing, but at an increase in bandwidth that is not justified in most facsimile services.

Two methods of obtaining this type of signal are in use. The first, shown in Fig. 9, consists essentially of a beat oscillator with one of the oscillators being shifted over a small percentage of its frequency by a reactance tube. As the light to the phototube increases, the frequency of the variable oscillator will be lowered. The two oscillators are set at sufficiently high frequencies so that the reactance tube can readily swing the required number of cycles without changing amplitude. For the swing of 1500 cycles on white to 2250 cycles on black the fixed oscillator could be set at 100,000 cycles and the variable one swung over the range from 102,250, with the phototube dark, to 101,500 cycles with the phototube receiving maximum light for white.

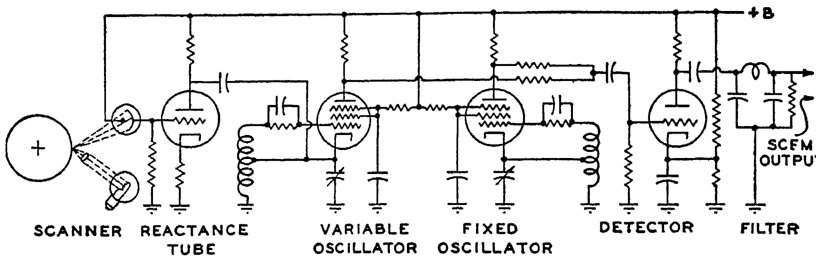


FIG. 9. Use of Reactance Tube and Beat Oscillator to Obtain Subcarrier Frequency-modulated (SCFM) Scanner Signals

A second method of obtaining SCFM signals directly, without heterodyning to obtain the low frequencies, is shown in Fig. 10. A resistance-capacitance oscillator of the 180° phase shift type is varied in frequency directly by using a tube control system as a variable resistor in one mesh of the phase shifting ladder network. When the control tube has zero input (phototube dark) the bias is adjusted to set the low-frequency end of the swing. As light to the phototube increases, the tube resistances decrease and the frequency of the oscillator is raised. Input volume from the phototube is adjusted so that the high-frequency end of the swing is just reached for white. When the proper network and tube constants are chosen, a linear range as high as 2 to 1 in frequency may be obtained with little change in amplitude.

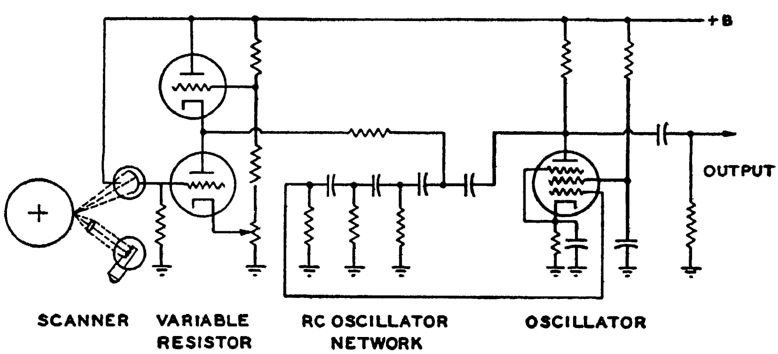


FIG. 10. Use of Tube Control on RC Oscillator to Obtain Subcarrier Frequency-modulated (SCFM) Scanner Signals

This circuit, when connected as shown, will give positive modulation, that is, an increase in frequency for an increase in light. To obtain negative modulation a reversing tube may be connected between the control grid and phototube, or the phototube may be reversed and connected anode to grid, and cathode to a negative supply potential below ground.

The changes in frequency with this circuit are very nearly instantaneous, because there is little stored energy in the network. At the same time, the frequency stability is adequate for the purpose.

It is sometimes necessary to use an existing scanning amplifier system having a-m output and still transmit signals of the SCFM type. A converter is then used in which the SCAM signals are rectified and filtered to obtain the original facsimile signals, and these are then applied to the control grid of an SCFM generator of either of the above types.

RECORDING SYSTEMS

A perfect facsimile recorder will build up a copy of the signals exactly as received, adding or subtracting nothing, and thus deliver a recording limited in detail only by the scanner and intervening transmitting circuit. The finished picture will be almost identical in appearance to the original copy.

Of the many recording methods, the four most generally used will be described here: photographic recording; wet electrolytic recording; dry electrolytic recording; and carbon-paper recording. Each of these systems has advantages possessed by none of the others and, therefore, will have particular uses to which it is the best adapted.

The length of the scanning line and the number of scanning lines per inch are generally the same as for the scanner, but this agreement is not necessary. The recorder copy may be made smaller or larger than the original by properly choosing the proportions of scanning-line length to line advance. The product of the total length of the scanning line and the number of lines per inch is called the index of cooperation; if this value is held constant, any size recording may be made with all dimensions correctly proportioned to those of the original copy. Thus, if the scanner has a total line length of 9 in. and is transmitting at 100 lines per inch, the index of cooperation would be 9 x 100 = 900. If it is desired to receive this picture on a recorder having a scanning-line length of only 4.5 in., the line advance would be made 200 lines per inch, and the received copy would be exactly one-half size.

Other than this index of cooperation, the other essential factor is the number of lines transmitted per minute. This is (often) termed "strokes," and the "per minute," which should be added, is understood. Thus the numbers 40–900 would signify a facsimile picture having an index of cooperation of 900 and transmitted at the rate of 40 strokes per minute.

4. PHOTOGRAPHIC RECORDING

In recording photographically, the sensitized paper or film is generally wrapped on the surface of a drum and is scanned by a small spot of light. The light spot is varied in intensity or size to record the different values of picture density. It may be varied in several ways—electrically, mechanically, or by means of polarization. In the electrically varied light, a neon or other gas discharge lamp is modulated in intensity by the signals. With the mechanical system, the light is steady, and varied either in intensity or size of spot by means of a vibrating shutter or diaphragm. In the polarized system, the Kerr cell is interposed between the light source and the picture drum, and the light is polarized before reaching the cell. The angle of polarization of the cell is changed by the picture signals, allowing more or less light to reach the picture.

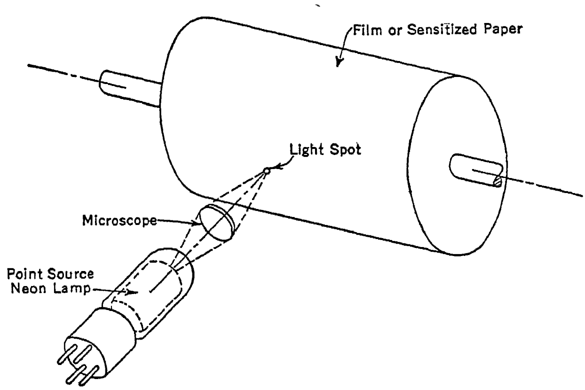


FIG. 1. Photographic Recorder, Using a Neon Lamp

The first method is more generally used in this country, and a simple recorder of this type is shown in Fig. 1. Here the lamp is of the "point-source" type. An intense illumination is produced in a small aperture within the lamp itself, and an image of this aperture is projected onto the surface of the drum by a lens system. The spacings of the lamp and lens system are so arranged that the aperture image is exactly the width of a scanning line. If several values of line advance are to be used with the same optical system, a variable diaphragm is introduced to regulate the size of the image to the proper value for the line width desired.

The relative motions of the optical system and drum may be any of those used in the simple drum scanner. Usually the drum is rotated while the optical system is gradually advanced along its surface.

The second method of photographic recording involves "valving" the amount of light reaching the paper from a steady light source, usually a tungsten-filament lamp. This can be done by placing an oscillograph mirror and aperture in the light path, the position of the mirror being varied electrically to change the area of the aperture exposed. As the amount of light will be directly proportional to area, a smooth variation of light with signal is obtained. Another method consists of placing a thin ribbon in the light path, just closing an aperture. As the ribbon is twisted by the incoming signals, light is allowed to pass through the aperture on both sides of the ribbon. This system produces a "variable width line" type of recording similar in appearance to a zinc etching, or, if the aperture is placed at 90° to the scanning direction, it will produce lines of constant width but variable density.

In the third method of using polarized light, a Kerr cell is utilized to change the light intensity. The optical system consists of two Nicol prisms placed between the light source and the aperture. These prisms are polarized in the same plane and therefore pass light through the system. The Kerr cell is interposed between the prisms, and applying signals to its polarizing plates will change its light-polarizing properties. The amount of light leaving the system is therefore controlled, and a true modulation of the light may be obtained. This system has been used for a number of years in Europe.

The photographic system is far the most accurate in its ability to reproduce completely the signals received and therefore is used in almost all commercial picture circuits. It has one serious disadvantage, however, in that the received picture must be developed before the results are known. The machine must be loaded and operated in the dark. In a fast service this developing is quite a handicap, and the fact that the picture cannot be seen until developed allows possible errors in the setting of the equipment to go unnoticed until the full time of transmission and developing has elapsed.

Most picture circuits are operated at speeds of 6 to 10 square inches per minute, this low speed usually being due to circuit limitations. With adequate light and sensitive films, the photographic recorder is capable of speeds far in excess of this value.

5. WET ELECTROLYTIC RECORDING

Electrolytic recording is similar to photographic recording in chemical action but has the advantage of being visible at once, or almost at once. It may or may not require some form of processing to make the recording permanent, depending on the chemicals used.

The principle of operation is that certain chemicals turn very dark when an electric current is passed through them. If a paper is saturated with such a chemical and scanned by a stylus contact, it may be darkened by current at each signal for black and thus build up the facsimile picture.

The common solutions are organic dyes, though silver or iron salts have sometimes been used as in photography or blueprinting. Some of these solutions react very rapidly but require high current density to bring about a dense enough black; others react with much less current but require some form of washing or fixing to prevent fading.

One recording solution, which gives a dense black permanent recording, used a special steel printer bar that is gradually worn away in the recording process. This chemical process is capable of speeds up to 50 square inches per minute. The chemistry of the color formation is given in U. S. patent 2,358,839. Another type of recording solution, which uses a platinum bar that does not take part in the chemistry of recording, is given in U. S. patent 2,306,471. The dye formed in this case may be any one of several of the azo dye family, and is usually of a deep purple color. Speeds up to 160 square inches per minute have been obtained with this type of solution.

A machine using a stylus would be a simple drum scanner with a dragging contact point on the surface of the paper. Another form of electrolytic recorder requires a continuous roll of supply paper and prints one picture after another without reloading. One form of this continuous type of recorder is shown in Fig. 2.

Here the scanning is done by a combination of a printer bar and a helix on opposite sides of the paper. The raised helix rotates at the same speed as the scanner drum, thus making one complete turn in the length of a scanning line. The point of intersection of this helix and bar will therefore travel across the paper once for each scanning line. Current for printing is passed between the helix and bar, through the paper. The bar is sprung slightly and allowed to drag over the damp paper surface to secure good contact.

In this machine the paper must be moist to conduct the printer current and allow the chemical reaction to take place. For the particular machine described, the paper is impregnated with the chemicals and kept at the proper moisture content by storing in sealed cans. The recorder itself is of moisture-tight construction so that the moisture in the paper is retained until after printing.

In another form of electrolytic recorder a dry untreated paper is threaded through a trough containing the recording chemicals before being fed between the helix and printer bar. After printing it passes over a hot ironing roll to dry and smooth out the recording.

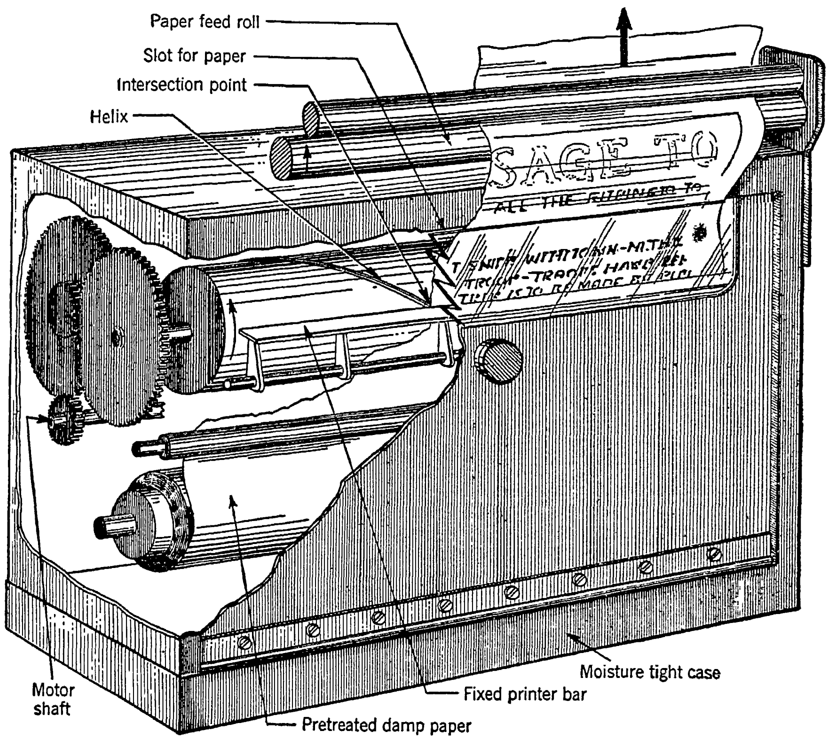


FIG. 2. Recorder Using Wet Electrolytic Paper

The advantages of this type of recorder are its simplicity, its visible recording feature, and the extremely high speeds of which it is capable. Recording speeds as high as 80 or 90 square inches per minute at 120 lines per inch definition are easily attained, allowing a full-sized page of 8 $^1/_2$ by 11 inches to be recorded in 1 minute. Though the handling of wet paper in the machine is awkward, this disadvantage is outweighed in many applications by the speed of recording.

6. DRY ELECTROLYTIC RECORDING

Several dry electrolytic recording papers have been developed for message service facsimile recording: they are much easier to handle than the wet electrolytic papers. though not capable of as high a recording speed. The best known of these, trade-named Teledeltos, has a light gray coating on a dense black paper base that has high electrical conductivity. When current is passed from a small stylus through the paper the metallic coating is burned away, leaving the black paper underneath exposed. The facsimile recording is built up by scanning with a stylus and partially or completely burning the coating where gray or black are to appear. See Fig. 3.

The advantages of this type of recording are that a permanent copy is produced with no processing, the dry paper is easy to handle, and the recorder, with only a stylus and drum,

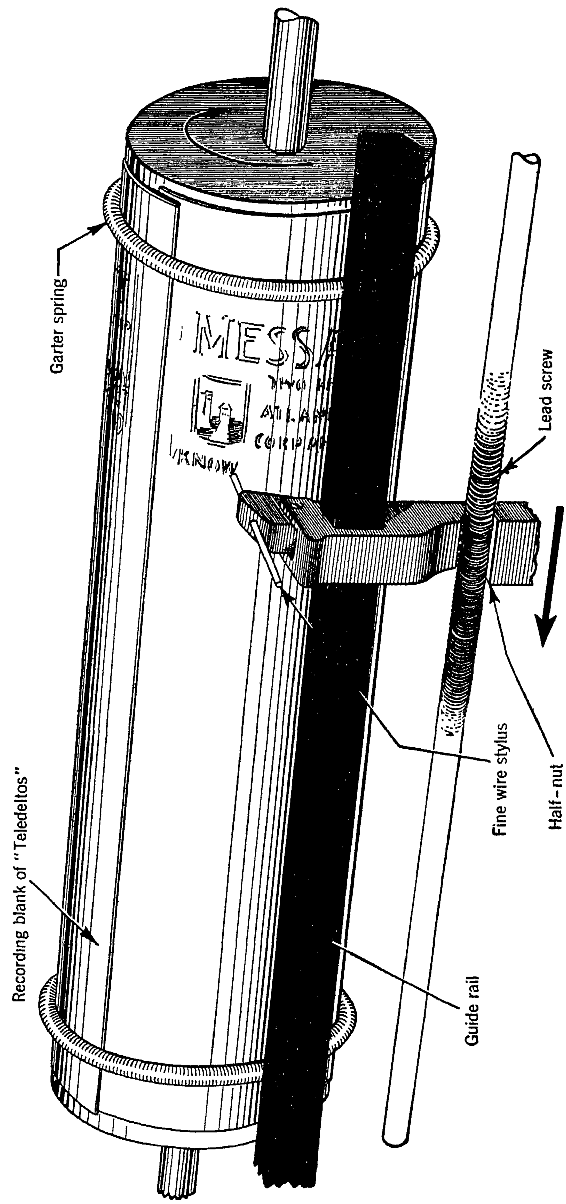


FIG. 3. Recorder Using Dry (Teledeltos) Paper

is mechanically simple. The disadvantages are that the contrast range of the finished copy is reduced by the gray background, and the halftone scale is not as linear as the photographic or electrolytic recorders. These do not greatly affect its value for message service where the subject matter is almost entirely made up of typing or handwriting.

7. CARBON-PAPER RECORDING

The first carbon recorder consisted of a stylus dragging over carbon and white papers wrapped on a drum. The stylus was moved down to give pressure for black, and lifted for white. This is a very simple form of recorder, but it has the disadvantage of the photographic recorder in that the picture is not visible until the drum is stopped and the carbon paper removed. It has the advantages of cheapness and simplicity, and the picture requires no processing to be permanent.

A later form of carbon recorder, illustrated in cross-section in Fig. 4, overcomes the disadvantage of invisible recording. Here the scanning is accomplished with a helix and printer bar, as in the continuous electrolytic recorder. Carbon and white paper are fed between the bar and helix, and after this are separated so that the surface of the white

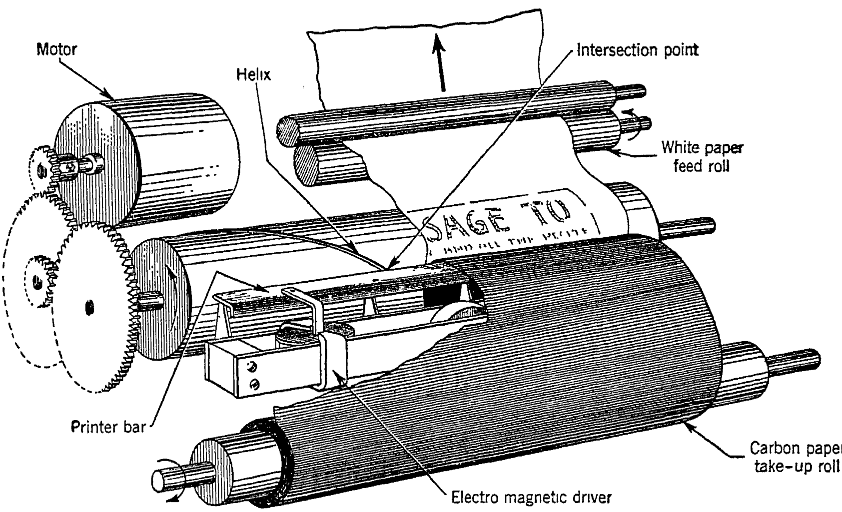


FIG. 4. Continuous-feed Carbon Recorder

paper is visible only a few seconds after the printing process. The bar is not allowed to drag the paper but is normally held away from it by an electromagnetic drive unit.

A signal for black depresses the bar, and a black dot is made by the pressure at the intersection of the bar and helix. The carbon paper is drawn over guides and wound up on a take-up spindle. The white paper is fed by a knurled feed roll with a series of rubber idlers held against it, similar to the paper feed of a typewriter.

Only one electromagnetic driver is shown for the printer bar. However, if wide paper is used, more than one driver may be necessary and the separate units will be equally spaced along the bar.

This method of recording is very simple, uses cheap paper, and prints a very good copy at speeds up to 10 square inches per minute. It is quite reliable, and the complete copy, with no processing necessary, is visible only a few seconds after recording. Its limitations are also pronounced. The printer bar is necessarily heavier than a stylus, and therefore the speed of recording is limited. Almost any carbon paper that may be used here will be soft enough to smudge a little when rubbed in the fingers, the same as a carbon copy from a typewriter. More mechanical accuracy is required in building this printing bar than in the electrolytic recorder, as the bar and helix must be parallel to within a few thousandths of an inch. The depressive motion of the bar is quite small, and, therefore, a little discrepancy in lining up the bar and helix will result in failure of part of the paper to be printed to a full black. Damping of the bar to eliminate "bouncing" and echo printing is somewhat of a problem, but it can be solved by over-powering the printing mechanism and absorbing the excess power in a damping arrangement on the bar itself.

One advantage mentioned separately here for emphasis is that this type of recorder may be used to print more than one copy at a time. If the printer bar action is made sufficiently powerful, several rolls of white and carbon paper may be threaded into the machine and a number of copies of the facsimile made at the same time. As many as 8 separate copies of a message have been made experimentally. Also the carbon paper may be of the "hectograph" type and extra copies of the recording may then be made by the usual duplication process of hectographing.

8. COMPARISON OF RECORDING METHODS

The recording method chosen for some particular service will depend on the quality of the copy required, speed of transmission, cost of the recording paper, ease of operation, and many other factors. In a news picture service, quality of the finished recording is most important, for the pictures are used as masters to make printing plates. Photographic recording is therefore used for all these pictures, the other factors being considered of less importance. Operating speeds are maintained as high as the wire lines or radio circuits will permit, generally 9 or 10 square inches per minute with 100 lines per inch detail.

For message services, speed of transmission is the most important, with simplified equipment able to run unattended as a next requirement. With somewhat limited bandwidths available over wire lines, the high speed of the wet electrolytic process cannot be attained, and the dry electrolytic process with its simpler recorder structure more readily fits the requirements. Operating speeds up to 30 or 40 square inches per minute can be maintained with this type of recording, if the transmission band of the line will permit. The more usual speed over most lines available for this type of service is 16 to 20 square inches per minute.

The requirements for broadcast facsimile of flash news services are primarily for medium speed, direct printing, and a minimum paper cost. Though the wet electrolytic recorder does not use the cheapest paper, it meets the other requirements on speed and direct printing. The copy appearance is pleasing, and the detail is adequate even at higher speeds than those now contemplated. Such broadcast services will probably be most effective at speeds of around 30 square inches per minute. Less than this is too slow, and higher speeds will tend to increase apparatus and paper costs too greatly.

Carbon recording uses the cheapest paper and can run unattended for long periods of time. However, it is the slowest of the recording methods, owing to the mechanical motion required of the printer bar. Its highest operating speed at present is about 10 square inches per minute, inadequate for most services but sufficient in some special instances where speed is less important.

9. RECORDING AMPLIFIERS

The signals received for facsimile recording will usually be in the form of an a-m or f-m tone. This may be either the original SCAM or SCFM tone transmitted over line or radio by the scanner, or it may be one obtained by heterodyning in a radio transmission of direct frequency modulation. If of the SCAM type, the signals will be applied directly to one of the recording amplifiers described later. If of the SCFM type the signals must first be changed into an equivalent SCAM signal by limiting and then passing the constant-amplitude signal through a slope demodulating filter.

Such a converter unit is shown in Fig. 5. The incoming SCFM signal presumably will have spurious amplitude modulation superimposed by fading if received by radio or

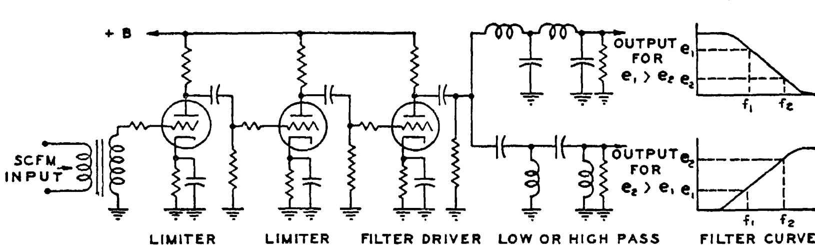


FIG. 5. Limiter Amplifier and Slope Demodulating Filters for SCFM-type Signals

changes in line transmission characteristics if received by wire line. Therefore it is first passed through a limiter amplifier of several stages, so that the limiter output signal will be of constant amplitude over wide changes in input level. This constant-amplitude signal then goes through either a low- or a high-pass filter, or both if pushpull output is desired.

These filters are designed to have slow, straight cut off slopes so that the amplitude of the output signals will vary linearly with frequency.

In the circuit shown the two end frequencies of the SCFM swing are labeled f_1 and f_2 for the white and black frequencies. If demodulated in the upper filter the output will be greatest for f_1 and least for f_2, while the output of the lower filter is the opposite. For undistorted demodulation these filter slopes must be linear beyond the band from f_1 to f_2 by

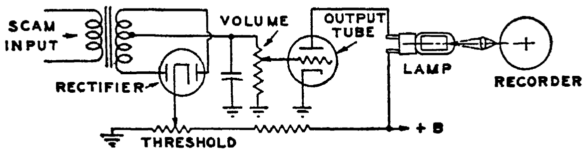


FIG. 6. Printer Amplifier for Maximum Output Current with Minimum Signal Amplitude

the amount of the side frequency spread of the SCFM wave. For the previous example of a swing of 1500 cycles for white to 2250 cycles for black, and band spread from 1125 to 2625 cycles, the filter slope must be linear over the entire range from 1125 to 2625 cycles.

The output of either slope filter will be the equivalent of an SCAM signal in amplitude envelope and can be used in the printer amplifier in the same manner.

Four simplified diagrams of printer amplifiers to actuate the various types of recorders are shown in Figs. 6, 7, 8, and 9. The first two are suitable for photographic recorders using a glow discharge lamp to expose the film or paper. In each the SCAM signal is rectified and

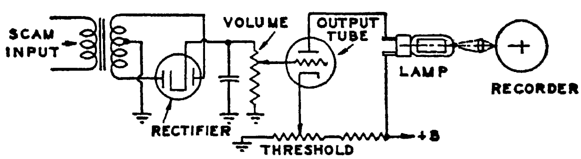


FIG. 7. Printer Amplifier for Maximum Output Current with Maximum Signal Amplitude

filtered by a capacitor across the volume control to obtain the facsimile signals. The output tube is then either driven to lower output current as signal amplitude increases, as in Fig. 6, or to higher output current as amplitude increases, as in Fig. 7. The particular one used will depend on the direction of modulation of the signal, and whether the drum is loaded with film to make a negative or with photographic paper to make a positive copy.

For recorders using a pushpull magnetically driven printer, or driver units of a carbon recorder, these two types of amplifiers may be combined, as in Fig. 8. As signal amplitude increases, the upper output tube will be driven to lower current and the lower tube to higher

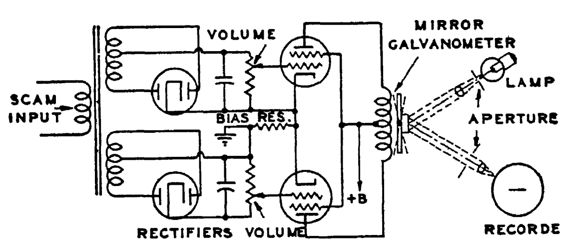


FIG. 8. Push-pull Printer Amplifier for Magnetically Driven Printer Systems

current. When properly adjusted true pushpull output is obtained, and the sum of the two output tube currents will be constant. This constant-sum current can be used as shown in the cathode return bias resistor to furnish the threshold bias required by the lower output tube.

For electrolytic recorders of either the wet or dry type it is a good safety measure to run the helix or recorder drum at ground potential. To accomplish this an amplifier such as in Fig. 9 may be used. A pushpull power amplifier at cut-off bias amplifies the SCAM signals to a power level sufficient for printing. This a-c signal is applied directly between the ground drum and stylus for dry electrolytic recording, but it must be rectified and applied between helix and bar in the proper polarity for most wet electrolytic printing processes. In some of these the color forms on the anode side, in which case the bar is made positive with respect to the helix; in others the color forms on the cathode side of the paper, and so the bar is made negative.

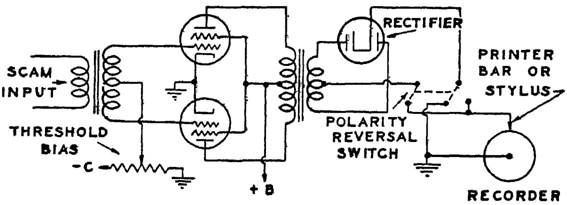


FIG. 9. Printer Amplifier for Either Polarity of Electrolytic-type Recording. With non-polarity sensitive papers, such as Teledeltos, the a-c output is used without rectification.

SYNCHRONIZING AND PHASING

10. SYNCHRONIZING

In every facsimile system, it is necessary that the recorder follow the scanner over the paper in order to produce an undistorted recording. The principle of synchronizing may be better understood by referring to Fig. 1. For clarity the picture elements are shown much larger in proportion than they really are. As the scanner starts the picture on element 1, the recorder also starts on its element 1. As succeeding scanning lines are drawn, the recorder must follow exactly, or the copy will be distorted by a misplacing of the elements.

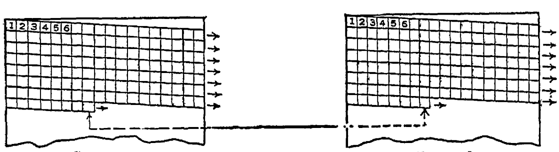
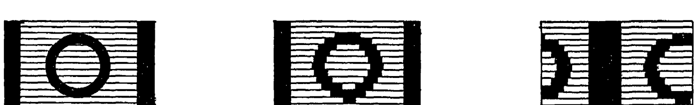


FIG. 1. Synchronism of a Facsimile System

Besides having the synchronizing correct, the recorder must be in "phase" with the scanner, as illustrated in Fig. 2. Even though the two drums are rotating at exactly the same speed, if they are not in phase the border of the picture will be misplaced. The recorder drum must start each scanning line at the same time the scanner is starting that scanning line, or, as shown in Fig. 2*C*, the border will be somewhere between the two edges of the paper instead of being exactly divided. Phasing and synchronizing become the same problem only if the phasing line of the picture, or border, controls the speed of the recorder. Where the synchronizing frequency is independent of the "phasing line," or of a much higher frequency than that of the phasing line, the two problems are separate and must be treated separately. This is generally the case in most commercial systems in use today.



A. Original in scanner *B.* Recorder in phase *C.* Recorder out of phase

FIG. 2. Phasing of a Facsimile Recorder

Before going into the means of synchronizing and phasing, the effect of imperfect synchronizing should be shown, to illustrate the problem better. Figure 3 shows the effect of an error in synchronizing on a unidirectional scanning system and on a back-and-forth scanner. The error illustrated here is that the recorder is running faster than the scanner by a very small percentage. In scanning the vertical line in *A*, the recorder gets farther along its scanning line each time, moving the recorded line farther and farther to the right.

The result, in a unidirectional system, is shown in *B*. In a back-and-forth scanning system, the result is much more pronounced, alternate lines moving apart, as in *C*. The result, for the recorder being too slow, would appear the same with this method of scanning, while with the unidirectional system the line would have slanted down to the left instead of down to the right.

The accuracy of the synchronizing may vary with the particular system. In commercial work, the necessary accuracy is very high. In a system scanning at 60 strokes, 100 lines per inch, and each line of 9-in. length, the total length of scanning line per vertical inch of paper is 900 in. In a picture of 10-in. length, this total scanning line length will then be 9000 in. A good copy will be made if the total drift in the border of the picture is not over $1/4$ in. in this 10-in. length of picture. Thus, the synchronizing system must hold an accuracy of $1/4$ part in 9000, or 1 part in 36,000. It must hold this rate for the whole transmitting time of nearly 17 minutes. Actually most commercial systems have synchronizing equipment accurate to 1 part in 100,000 or better.

A. Original in scanner *B*. Recorder fast, unidirectional scanning *C*. Recorder fast, "back-and-forth" scanning

Fig. 3. Recorder Not Perfectly Synchronized

TUNING-FORK FREQUENCY STANDARD. In short-distance facsimile transmission, as for instance local coverage of a broadcast facsimile service, the same a-c power supply is often available for both scanner and recorders. Synchronism is then simplified by driving both scanner and recorders with ordinary synchronous motors connected to the common supply. In long-distance transmission, or across the sea, this is not possible, and synchronism is generally maintained by controlling the motors of both scanner and recorder by accurate frequency standards. Such frequency standards usually take the form of very accurate tuning forks that will hold a constant frequency to within 1 part in 100,000 or better. Crystal standards could also be used, but, as the control frequency for the motor is usually low, less dividing of frequency is required with a fork.

To hold the required accuracy, the fork is usually held at a fixed temperature by thermostat control, or a temperature-compensated bimetallic type of fork is used. Figure 4 shows one method of driving a tuning fork by using it as a resonant coupling circuit in a vacuum-tube oscillator. By changing the driving power supplied to the fork, a vernier control on its frequency is obtained. The fork frequency may be amplified by tubes or thyratron inverters to a power level sufficient to drive a synchronous motor directly, or it may be used in other ways to control motor speed.

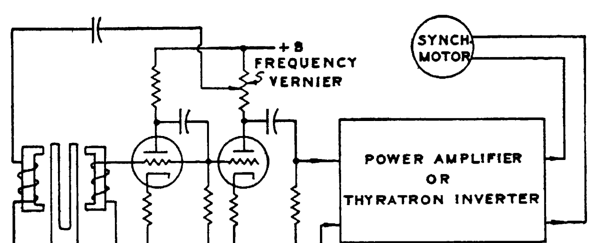


Fig. 4. Fork Control of Synchronous Motor

In one application, an 1800-cycle tuning fork is used both to supply carrier tone for an SCAM signal and is also amplified to a power of about 10 watts to drive an 1800-cycle synchronous motor. The synchronous motor is brought up to speed by a d-c motor, as it is not self starting. In applications where a standard 60-cycle synchronous motor is used, the fork frequency can be made 60 cycles, or divided down to 60 cycles from some higher frequency.

MAGNETIC BRAKE SYNCHRONIZING. Another type of control circuit is shown in Fig. 5. An induction motor, or other type of motor having good speed regulation, is used to drive the recorder or scanner and is controlled to an exact speed by means of a magnetic brake. A tone generator of the phonic wheel type is mounted on the motor shaft, having the correct number of poles to generate a frequency equal to the fork frequency at the correct motor speed. A phase comparison between this generated frequency and the fork control frequency is then used to vary the brake current and make the system lock into synchronism with the fork.

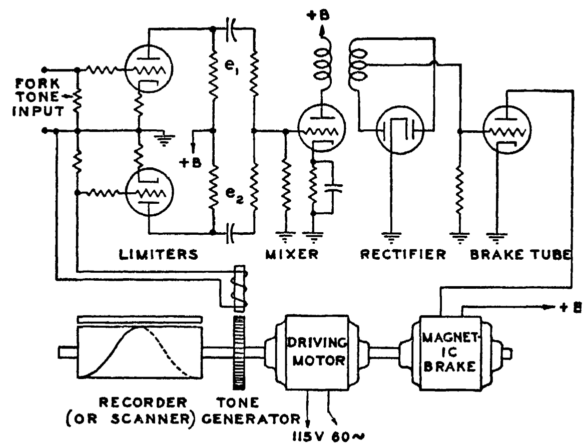


FIG. 5. Magnetic Brake Synchronizing System

A wave analysis of the brake action is shown in Fig. 6. The generated tone and the fork tone are each amplified and limited to give the two square waves e_1 and e_2, as in Fig. 6*a*. These two waves are added together algebraically in the mixer tube to give the waves shown in Fig. 6*b* and are full-wave rectified to give the waves in Fig. 6*c*. The result is then the grid voltage of the tubes supplying the braking current, shown in Fig. 6*d*.

Two conditions are illustrated. The waves on the left side show the motor leading the fork by a small phase angle, and the pulses of brake current are of full amplitude but narrow in time, so that the average brake current is small. If the motor tries to speed up for any reason, such as an increase in line voltage or lightening of the mechanical load, it will advance in phase with respect to the fork. The waves on the right side illustrate how the brake current pulses are increased in time width to give a higher average brake current that tends to slow down the motor. The brake is thus turned full on, or full off, by the square pulses, and the correct average is obtained mechanically rather than by smoothing these pulses in a filter. Hunting is thereby almost completely eliminated. As the two waves are squared up before comparing, the change in ratio of time on to time off is a linear function with phase-angle changes of zero to 180°.

As very effective brake action can be obtained by passing the pulsating direct current through the windings of an ordinary induction motor, a special design of brake is not necessarily required. This system is especially well adapted for very high-speed facsimile where motor powers as high as 1/6 hp are needed and control must be at high frequency to limit phase displacement with changes in load or line voltage.

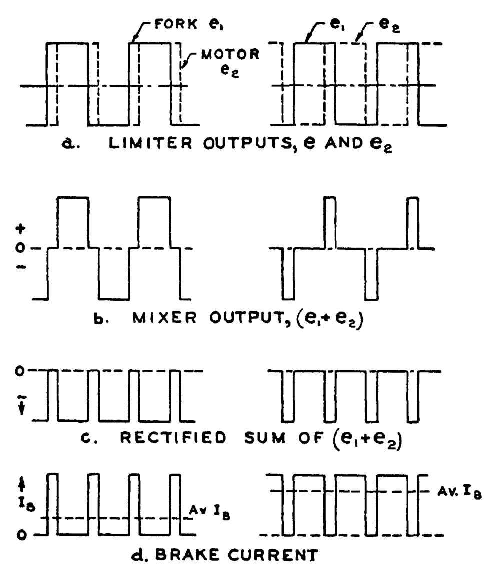


FIG. 6. Brake Operation as Motor Goes from Small Phase Lead on Fork to Large Lead

START-STOP SYNCHRONIZATION. The first methods used for facsimile synchronization were generally of the start-stop type, and, although such a synchronizing system is now practically obsolete for facsimile, it is still used on some forms of automatic tape printers, such as the teletype.

In start-stop systems, the scanner is generally operated at a constant speed and has a "phasing line" of a considerable time length. During this phasing line interval, the recorder will have finished its scanning line and stopped automatically. The scanner sends a pulse at the start of the succeeding scanning line, and a clutch, or similar mechanical apparatus, starts the recorder on the next scanning line. A governor-controlled motor, or some other fairly accurate drive, is used to maintain the recorder at a constant speed for the duration of each scanning line.

The chief merit of this system is that the errors in speed of the recorder are not accumulated, each scanning line starting afresh. The greatest possible discrepancy in synchronizing, therefore, is the error in any one scanning line itself, and this can be made quite small. The disadvantage is that the mechanics of such a system must be quite complicated, and a definite starting pulse must be received or the entire scanning line is lost. The speed of the entire system must, therefore, be quite slow to insure that these two factors do not interfere with the picture. A complicated scanning system cannot be started instantaneously at a high scanning speed, as allowances must be made for the inertia. Fading of the signal, if received by radio, would cause such a system completely to miss whole scanning lines if starting pulses were not received.

For use by line, such a system has advantages, as an ordinary governor will synchronize a motor accurately enough for the purpose, and failure to receive a starting pulse is rare.

OTHER SYNCHRONIZING SYSTEMS. Certain recording systems require no synchronizing at all, and such methods, sometimes used for cable transmission of pictures, involve setting up a certain number of picture elements by machine, or by hand, and sending a tape of this series of elements in numerical order. The recording is then assembled by hand, usually requiring a competent artist to give the picture a lifelike appearance. This method has been used for a number of years with great success over wire and cable. The Bartholemew-McFarlane system or, in shorter terms, the "Bartlane" system is a variation of this method.

The synchronizing frequency of the scanner is sometimes sent over the radio or wire line, and an amplifier is used to build this signal up to a value where it is able to drive or control a synchronous motor on the recorder. Such methods are satisfactory on line transmissions and short radio circuits but cannot be depended on for long radio transmissions.

11. PHASING

The phasing of the recorder to the incoming signals can be accomplished either manually or automatically. The simplest manual method is to throw the recorder out of synchronism and let it drift until some indicator, such as a neon lamp fed by the phase signal pulses, indicates in phase, and then to re-establish synchronism to hold this position.

In a simple automatic system, used on many news photo equipments, the recording drum is driven through a clutch, and a projecting ear on the drum is arranged to engage against a stop-pin having a magnetic release. At the start of the picture the drum is held in start position by this pin, and the clutch slips. When the start phase signal is received the drum is released and, being of low mechanical inertia, starts rotating almost immediately and in phase. The stop-pin is locked out automatically when tripped, so the drum continues to rotate for the duration of the picture transmission time. A phasing system suitable for a continuous recorder is shown in Fig. 7. A phase signal is transmitted at the start of each

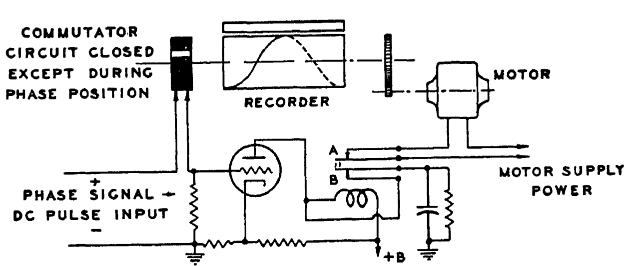


Fig. 7. Automatic Phasing of Continuous Recorder

scanning line, and this signal is selected out of the picture signals by frequency or amplitude discrimination.

It is rectified and passed to the input circuit as a d-c pulse at the start of each line. In series with this input is a commutator that is closed at all times except for a short gap at the correct phase position. When the recorder is running in phase, this gap opens the circuit for a slightly longer time than the duration of the phase pulse, so the tube receives no signal. If the pulse arrives at any other time, it passes through the closed portion of the commutator and causes the tube to draw a pulse of current to operate the relay. This opens the motor circuit momentarily and causes it to drop synchronism. To insure the relays staying open long enough for the motor to lose 1/2 or 1 cycle of synchronism, the contacts *B* of the relay connect the capacitor from relay coil to ground, and the charge current of the capacitor holds the relay in operating position for a fixed length of time. The resistor across this capacitor is too high to pass enough current to hold the relay in, but it bleeds the capacitor to zero charge between pulses. The motor is thus jogged out of synchronism once for each pulse received in an out-of-phase position, and this process continues until the correct phase position is reached, and the commutator again opens the pulse circuit at the correct time.

TRANSMISSION CHARACTERISTICS

12. WIRE LINE TRANSMISSION

Transmission of facsimile signals by either wire line or radio puts more exacting requirements on the circuit than telephone or telegraph transmission. This is largely due to the exact timing of the signals, which requires that delay equalization of a wire line be much more precise than for telephone work. Any appreciable difference in arrival time at the recorder of the high- and low-frequency components of the picture will show as transients and ghosts that exaggerate the outlines of objects in the picture and may even make typing unreadable. Accurate delay equalization is therefore required on all but short lines.

The amount of delay equalizing necessary is also directly affected by the speed of the transmission and by the bandwidth required. For the previous example of a maximum keying speed of 750 cycles, the shortest dot to be transmitted is 1/1500 second, or 0.667 millisecond. Any difference in delay equalization over the band of 1800 $\pm$ 750 cycles, or from 1050 to 2550 cycles, should not exceed a fraction of this 0.667 millisecond or noticeable distortion will result. In this case, the line should be delay-equalized to $\pm$ 0.25 millisecond over the 1050- to 2550-cycle bandwidth. In many long lines used for facsimile, this maximum delay error of $\pm$ 0.25 millisecond, over a band of 1000 to 2600 cycles, is maintained.

If the speed of transmission were doubled, the maximum permissible delay error would be halved and the bandwidth doubled at the same time. The problem of getting good enough lines is therefore increasingly difficult as speed is increased.

This exactness of delay equalization can be compared to regular voice circuits where 10 or even more milliseconds' delay difference does not appreciably affect quality of speech.

Where a-m signals are being used, the line must also be equalized for amplitude over the transmission band, but this is usually an easier problem. Where SCFM signals are used the amplitude characteristic of the line is relatively unimportant, but the delay characteristic must be as good as for SCAM signals.

For short transmissions of less than 100 miles portable scanners are sometimes operated into ordinary coin-box phones and over regular long-distance lines. This practice is satisfactory in some cases, but distortion is much greater than over the specially equalized lines, and picture quality is therefore lower.

13. RADIO TRANSMISSION

Radio transmission is beset with more difficulties than line transmission and, for long distances, is generally slower. Many factors enter in long-distance radio transmissions, such as fading, multipath delays, interference, echos, and other forms of distortion that must be corrected for, or else the speed must be decreased until the particular distortion present is reduced sufficiently to be no longer objectionable.

Rapid changes in transmission distance, due to varying heights of the ionized layers, give the effect of varying delay times on wire lines and are the limiting factor on speed of transmission. Some of the most pronounced effects can be eliminated by suitable directive antennas with limited pick-up angles, to eliminate the next higher order of skip or hop. When such antennas are used, and the proper choice of carrier frequency made for the distance, speeds up to 10 sq in. per minute are generally possible on circuits as long as New York to London, and a speed of 6 sq in. per minute is very reliable.

In earlier radio facsimile systems, dot-halftoning was developed so that a keyed on-off CW type of transmission might be used. Limiting the incoming signal then allowed most of the results of fading to be removed. However, all present long-distance transmissions are made either by using SCFM on regular voice transmitters, or direct frequency modulation, or frequency shift of the radio carrier. The dot-halftone systems have therefore become obsolete, and at the same time speeds have increased from 2 to 3 times that possible with the dot systems. With either of these newer methods limiting can be done either at radio frequency for frequency modulation or at audio frequency for SCFM and fading can be largely removed.

For short distances, or when using ultra-high-frequency relaying, the speed is not so limited, for multipath troubles do not enter, and wide bands may be used.

SPECIALIZED APPLICATIONS

14. DUPLICATORS

Many applications of facsimile have been made that illustrate that it is not limited solely to the transmission of pictures over long distances. The scanner and recorder can be mounted on the same shaft and a duplicator, or copying machine, obtained. Two types of facsimile duplicators are in use, each having definite advantages over other forms of duplicators in certain applications.

In one type of duplicator a wet electrolytic recorder is combined with a rapid-loading type of scanner, such as shown in either Fig. 4, p. 5, or Fig. 5, p. 6. The speed of operation is very high, 85 sq in. per minute, and a full-sized letter page 8 1/2 by 11 in. is copied in slightly over a minute, with 120 lines per inch detail. The wet paper passes over an ironing roll after printing, and a finished dry copy is thus delivered.

As both scanner and recorder are rigidly coupled to the same driving motor, no synchronizing or phasing is required. The amplifier system from phototube to printer becomes very simple, as there is no transmission and reception problem over line or radio.

This type of duplicator is useful where only a few copies each of a large number of originals are required. As reflected light is used the original may be opaque, entirely unsuitable for blueprinting. It thus compares with photo-copying (though the cost of the printing paper is less) but eliminates the necessity of developing and fixing.

MULTIFAX. In *Multifax*, a master mimeograph stencil is cut by facsimile methods so that a large number of copies of the original may be made. Copy containing illustrations and diagrams that would be almost impossible to make up by ordinary means can thus be obtained in quantity, without going through more expensive printing processes.

The scanner and recorder are on the same shaft, as in *Duplifax*, to eliminate the need of synchronizing and phasing. The stencil is clamped on a recording drum and cut by a stylus which is vibrated at high frequency. While vibrating, the stylus is moved towards the drum for black and delivers a large number of blows to the stencil to displace the wax. The stylus is retracted for white so that it just misses touching the stencil. The vibrating stylus has no tendency to drag out or tear the stencil, as it would if pressure only were applied, and so much finer detail can be realized than with hand-cut stencils.

A full letter-sized master stencil can be prepared from the original in 10 or 15 minutes by this method, about the time required for typing a stencil without illustrations. Another advantage is that the original is prepared on white paper, and printed diagrams or illustrations may be pasted in place without being hand drawn on the stencil.

15. TAPE FACSIMILE

Tape facsimile systems are a special adaptation of facsimile in which the recording is printed on a narrow slip similar to that used in news tickers. The scanning may be optical or mechanical, but the recorders for either type of scanning are usually of the helix and printer-bar type, somewhat like a miniature version of the carbon recorder.

Where optical scanning is used, the message is typed, or handwritten, on the transmitting tape, and this original copy then fed through a scanner similar to that shown in Fig. 1. A spot of light is traversed across the width of the tape by the combination of a rotating prism and a fixed system of three cylindrical lenses and one right-angle prism. One stroke across the tape is obtained for each face of the prism that passes. These scanning lines are very short, usually 1/4 to 1/2 in. The phototube signals generated by the reflected light are then used to obtain either SCAM or SCFM signals in the same manner as in any page facsimile scanner amplifier.

For mechanical scanning the transmitter takes a form similar to a tape-teletype machine, in which each letter is represented by a disk similar to an "Omnigraph" disk. Each disk is a commutator and keys a series of signals that will form a facsimile image of that character when printed by the recorder. Operation of this form of tape scanner is then similar

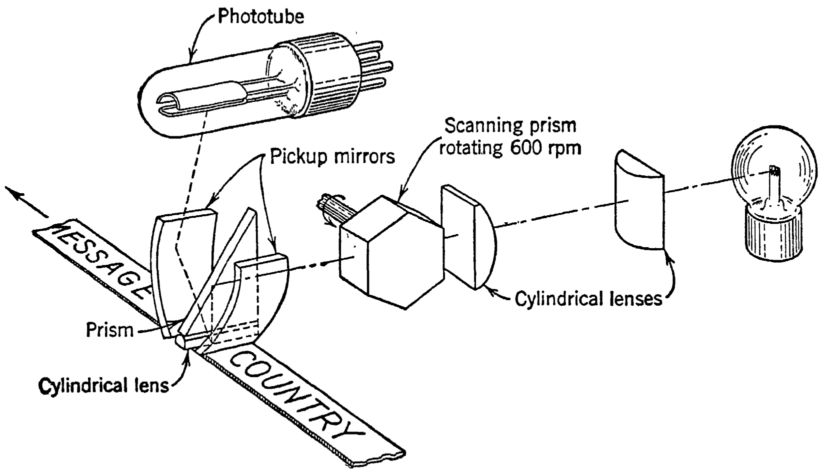


FIG. 1. Tape Facsimile Scanner

to that of a teletype machine, but with the output signals coded for facsimile recording rather than for operating a typewriter type of receiver.

Either type of scanner will operate the tape recorder shown diagrammatically in Fig. 2. The helix is very small for a scanning line as narrow as this, and it can be inked directly by the ink roller shown instead of by using a slip of carbon paper to supply the coloring matter as in the page type of recorders. Otherwise the recorder action is exactly the same as for the page carbon recorder shown in Fig. 4, p. 15. Tape systems are built for message service

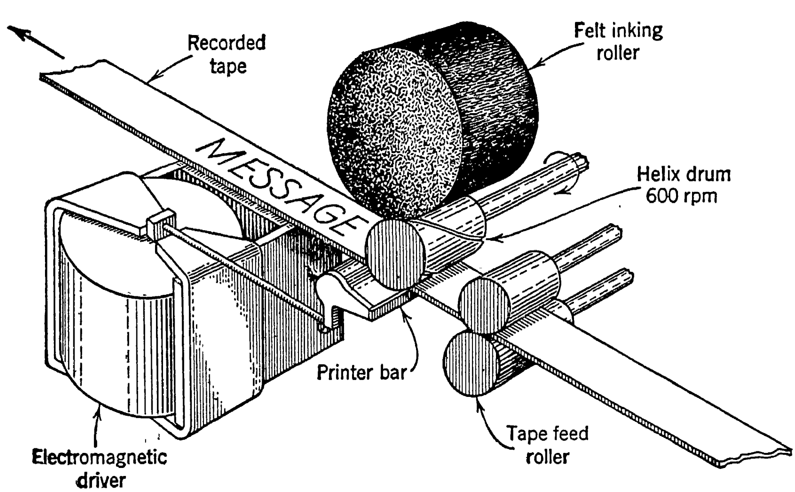


FIG. 2. Tape Facsimile Recorder

and do not need to have the fine detail usually required in a picture system. Whether optical or mechanical scanning is used, the characters transmitted are large block type to reduce keying frequency and to get as large a number of words per minute as possible in a narrow channel.

The scanning-line length is very little greater than the height of a letter, so that there is little waste time in margins. For instance, if $^3/_{16}$-in. block type is used, the scanning line will be about $^1/_4$ in. long. To transmit this type, the shortest dot necessary would be about 0.020 in., and a maximum number of cycles per scanning line would be 6. With the constants of 60 lines per inch and 60 lines per second, 1 in. of tape per second is transmitted with a keying speed of not over 360 cycles per second. This tape speed of 1 in. per sec will represent about 60 words per minute.

While this band of 360 cycles for 60 words per minute is about 8 $^1/_2$ times that required for the commercial automatic 7 unit codes, where 42 cycles (or bauds) per second represents 60 words per minute, this disadvantage is largely offset in instances where signals are poor, or noise high. With tape facsimile, interference can obliterate a letter, but it cannot make it print a wrong one.

BIBLIOGRAPHY

Standards on Facsimile, Definition of Terms, 1942, Supplement of *Proc. I.R.E.*, Vol. 30, No. 7, Part IV.
Radio Facsimile, RCA Institutes Technical Press, 1938, pp. 112 to 128 contains a very complete bibliography of facsimile up to October 1938.
Mathes and Whitaker, Radio Facsimile by Subcarrier Frequency Modulation, *RCA Rev.*, October 1939, pp. 131–154.
Artzt, M., Frequency Modulation of RC Oscillators, *Proc. I.R.E.*, July 1944, pp. 409–414.
Bliss, W. H., Subcarrier Frequency Modulation, *Proc. I.R.E.*, August 1943, pp. 419–423.
Felch, E. P., Measuring Delay on Picture Transmission Circuits, *Bell Lab. Rec.*, January 1936, pp. 154–157.
Hinshaw, F. A., Delay Equalizers for Telephotograph Transmission, *Bell Lab. Rec.*, February 1936, pp. 193–197.
Mertz, P., The Telephotograph Line, *Bell Lab. Rec.*, February 1936, pp. 178–184.
Mertz and Pfleger, Irregularities in Broad-Band Wire Transmission Circuits, *Bell Syst. Tech. J.*, Vol. 16, pp. 541–559 (October 1937).
Schulman, D., Facsimile Synchronizing Methods, *Electronics*, March 1946, pp. 131–133.
Wise and Coggeshall, The Handling of Telegrams in Facsimile, *Proc. I.R.E.*, May 1941, pp. 237–242.
U. S. Patent 2,358,839, Iron Bar Wet Electrolytic Recording.
U. S. Patent 2,306,471, Axo Dye Wet Electrolytic Recording.

SECTION 20

TELEVISION

PRINCIPLES AND THEORY

By A. V. Loughren

TELEVISION BROADCASTING

By T. J. Buzalski, A. L. Hammerschmidt, and F. J. Somers

TELEVISION RECEIVERS

By W. F. Bailey and R. J. Brunn

OTHER FORMS OF TELEVISION

By A. V. Loughren

TELEVISION

PRINCIPLES AND THEORY

By A. V. Loughren

Television is defined as "the electrical transmission and reception of transient visual images." *

Television technique for monocular, monochrome pictures has developed sufficiently to lead to adoption by the Federal Communications Commission of standards for broadcasting.

1. PHYSIOLOGICAL REQUIREMENTS

The performance required of a television system is determined by physiological requirements which must be met for the performance to be acceptable. Section 14 discusses these in detail. They vary from individual to individual, but generally acceptable design values for the several quantities have been arrived at based on extensive tests. These requirements include:

Resolution. An observer with good eyesight is able to resolve successive contrasting objects individually subtending as little as 1 minute of arc. A square subtending 1 minute of arc on a side corresponds to a solid angle of approximately 10^{-7} steradian. (References 1, 2, 3.)

Field of View. A normal eye is capable instantaneously of critically observing a field of the order of 0.001 steradian. Since the eye direction can be quickly and readily changed, a much greater field than this is available within a very short interval of time. For sustained viewing of images the viewing distance of four to eight times the picture height chosen by most observers produces an image field of the order of 0.02 to 0.07 steradian. Such a field is 200,000 to 700,000 times the minimum resolvable solid angle (reference 3).

Sharpness. Sharpness is the subjective quantity corresponding to the objective quantity "resolution." Figure 1 shows the relation between sharpness and resolution; it in-

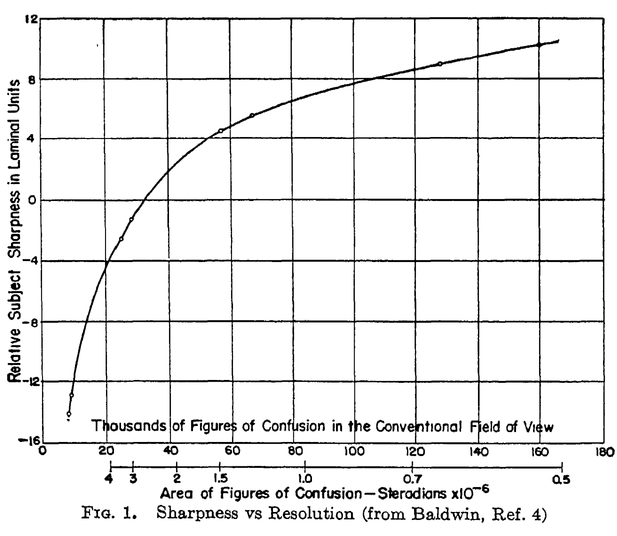


FIG. 1. Sharpness vs Resolution (from Baldwin, Ref. 4)

dicates that increasing the resolution by making the size of the figure of confusion less than 1.5×10^{-6} steradian increases the sharpness only slightly (reference 4).

---

* RMA.

Brightness. Because of its essentially logarithmic response and its ability to control admitted light by means of the iris opening, the human eye is capable of observing objects whose brightnesses lie within the range from 4×10^{-5} to 4000 ft-lamberts. It is found, however, that satisfactory viewing requires restriction of this range. Under conditions of low ambient illumination, highlight brightnesses as low as 1 ft-lambert are found acceptable; however, under conditions of normal artificial and natural lighting indoors, highlight brightnesses as great as 200 ft-lamberts are desirable. Values of 10 to 100 ft-lamberts are suitable for design purposes.

Contrast. The total contrast range instantaneously perceptible to the eye is believed to be about 40,000 : 1. However, reproductions exhibit contrast ranges from 10 : 1 for rather unsatisfactory images to 200 : 1 for the best photographic transparencies. Television pictures having a contrast range of 30 : 1 have been judged reasonably satisfactory.

Color. See Section 14 and references 5 and 25.

Depth. See Section 14.

Moving Objects. Ideally, reproduction of a picture of a moving object requires that each elementary area of the picture change synchronously with the corresponding changes in the original scene caused by the motion of the object. It is known, however, that the resolution of the eye for moving objects is much poorer than for stationary objects. It is consequently permissible to reproduce the picture at finite intervals rather than continuously. For most purposes the interval of $^1/_{24}$ sec is short enough to leave with the observer the illusion that motion is continuous rather than discontinuous.

Shape and Size of Picture. A rectangular shape with the width equal to four-thirds of the height has been found generally acceptable. This ratio is defined as the aspect ratio.

The minimum acceptable size for reproduced pictures is believed to lie in the range between 4 by 5.33 in. and 7 $^1/_2$ by 10 in. Smaller pictures produce fatigue within a short time unless special devices are worn by the viewer. The maximum acceptable picture size is determined primarily by the viewing distance available. For household use pictures up to 15 by 20 in. are suitable, while for use in halls and theaters much larger ones are appropriate.

2. SUBDIVISION OF PICTURE; EFFECT OF SCANNING RATES

Methods of transmitting and reproducing a television picture fall into two classes. Both classes depend on the subdivision of the scene into a sufficiently large number of elementary areas. In one class of system a separate transmission channel is provided continuously for each element. The large number of elements required for a satisfactory picture has shown systems of this class to be impracticable. In the other class, which all present practical systems use, the elements are connected to a single transmission channel successively in an ordered sequence common to both transmitter and receiver. This process is called *scanning*.

SCANNING. An area to be scanned is subdivided into elements each of which is connected to the transmission channel periodically in some regular sequence. In its simplest form the operation consists in starting at the upper left-hand corner of the picture, traversing successively the row of elements along the top of the picture from left to right, following with a similar traverse one element width lower, and continuing this process until the bottom of the scanned area is reached. A single traverse across the picture in one direction is called a scanning line. The complete scanning pattern is referred to as a raster. This process is shown diagrammatically, for a picture containing only a few scanning lines, in Fig. 2. The electrical frequency components produced by thus scanning a fixed image may be shown to consist of: (1) a d-c component; (2) components at the vertical scanning frequency and its harmonics; (3) components at the horizontal scanning frequency and its harmonics; (4) components at sum and difference frequencies of the above. Physically the d-c component represents the average brightness of the image. The components at the vertical frequency and its harmonics represent bands extending horizontally across the picture. The components at the line frequency and its harmonics represent vertical bands. The components at sum and difference frequencies represent inclined bands. If the picture changes with time, side bands are added to some or all of these components (references 6 and 29).

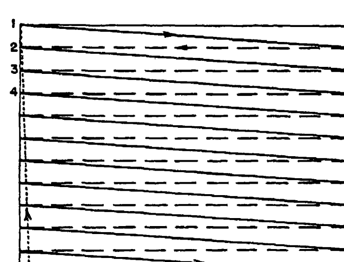


FIG. 2. Simple Raster

Vertical Resolution. The scanning spot in the usual case is not rectangular nor does it exhibit uniform effectiveness over its area. As a typical example of this, Fig. 3 shows the distribution of light intensity over the scanning spot of a cathode-ray tube. If the spot moves rapidly in one direction, forming a line, its effective distribution in the other direction assumes some such form as that shown in Fig. 4 at A. This figure also illustrates, at B, the condition under which a flat field of illumination is produced by the overlap of successive lines, and at C a line-to-line spacing which fails to produce a flat field. Similar relations exist with respect to the scanning spot in the photosensitive device at the transmitter.

The photo device at the transmitter should be adjusted to satisfy the flat-field criterion. Failure to do this results in a type of distortion known as "beads," which is not susceptible of any subsequent correction.

Vertical width of confusion, defined as the average width in the reproduced image of a very narrow line appearing before the transmitter, positioned at a slight angle with respect to the scanning lines, is equal to $\sqrt{2}$ times the scanning-line pitch (reference 7).

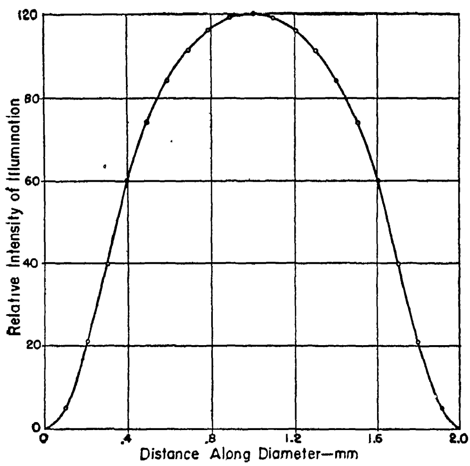


FIG. 3. Distribution of Intensity across Scanning Spot (After Zworykin, *Proc. I.R.E.*, Dec., 1933)

Horizontal Resolution. If the scanning spot of a pick-up device moves horizontally across a vertical line of negligible width, the resulting electrical impulse describes the horizontal characteristic of the spot. Typical forms of this impulse are shown in Fig. 5. The width of a rectangle having the same maximum height, and including the same area as the spot characteristic (as shown dotted in Fig. 5*b*), is an approximate measure of the duration of the impulse. Since the spot is traveling at a fixed velocity, the abscissa in Fig. 5 represents not only time but also distance. It may, for analysis, be transformed into steady-state amplitude and phase characteristics as functions of frequency (for example, by means of the Fourier integral theorem). The effective band width may be expressed approximately by the width of a rectangular area having the same maximum height and same included area as the frequency characteristic. It may be shown that to a useful approximation the effective spot width in seconds (or microseconds) is related to the effective band width in cycles (or megacycles) by the equation:

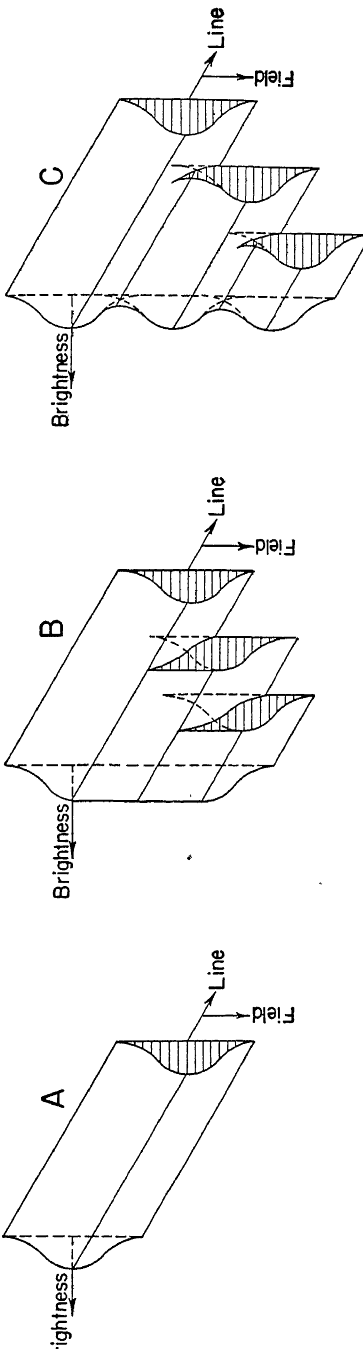
$$t = \frac{1}{2f_c} \tag{1}$$

where t is the time-duration of the equivalent rectangular electrical transient and f_c is the cutoff frequency of the equivalent rectangular frequency characteristic. A signal generated by a scanning spot moving across a narrow line as just described, transmitted electrically and reproduced by a reproducing device, will have its frequency spectrum modified both by the electrical circuits and by the equivalent transmission characteristic of the spot of the reproducing device. These several characteristics may be multiplied together to produce the overall transmission characteristic of the system. The corresponding effective horizontal width of confusion is then obtained by applying to this characteristic the inverse transformation by means of the Fourier integral theorem. As an approximation, the effects of several sources of limitation on the frequency band width, connected in cascade, on the effective overall band width, and on the duration of the shortest reproducible impulse are given by the equations

$$f_c = \frac{1}{\sqrt{\dfrac{1}{f_{c1}^2} + \dfrac{1}{f_{c2}^2} + \cdots}} \tag{2}$$

and

$$t = \sqrt{t_1^2 + t_2^2 + \cdots} \tag{3}$$



A. Single Line.

B. Several Lines Overlapping to Produce a Flat Field.

C. Several Lines Spaced to Produce a Wavy Field.

FIG. 4. Lateral Distribution of Scanning Line Brightness

So long as the frequency characteristic or the transient impulse form reasonably well approximates the error function

$$y = e^{-x^2}$$

the relations given above are close approximations (references 7 and 8).

The scanning spot is usually symmetrical and thus exhibits no phase distortion. The electrical circuits, however, are potential sources of phase distortion. Phase distortion affecting low-frequency components of the reproduced picture tends to alter the vertical or lateral shading of the picture. Phase distortion affecting high-frequency components tends to give the picture a "bas relief" effect in which edges of objects in the image may be preceded or followed by bright or dark outlines. Sharp cutoff in the amplitude characteristic produces "overshoots" superficially similar to high-frequency phase distortion. The two effects differ in that if a symmetrical object is scanned the distortion due to sharp amplitude cutoff will be itself symmetrical, whereas that due to phase distortion will be opposite in its character on the two sides of the image (reference 9).

FLICKER. Flicker is not inherent in television. It is a consequence of the use of scanning in conjunction with picture display devices in which energy is supplied to a given elementary area of the display device for only a minute fraction of the total time. The use of such display devices with the scanning sequence of Fig. 2. at a picture repetition rate of 25 or 30 per sec, produces severe flicker at brightnesses even below 1 ft-lambert. Restriction of the maximum brightness to this level is not acceptable.

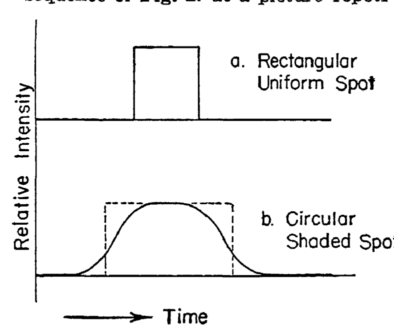


FIG. 5. Electrical Impulse Produced by Scanning a Narrow Vertical Line

A major improvement is obtained by interlaced scanning. In this form of scanning, shown in Fig. 6, the vertical component of velocity of the scanning spot is doubled as compared to that of Fig. 2 so that while the spot crosses the picture from left to right it falls an interval equal to *twice* the distance between scanning lines. If, therefore, the spot starts at the top center to trace line 1, after completing line 1 it goes to the left-hand side of the picture and starts line 3, continuing in this manner until it reaches the bottom right side of the picture. At this point the vertical retrace takes place and the spot returns to the top left of the picture and starts to scan line 2, followed by lines 4, 6, etc., until the lower center of the picture is reached. (For purposes of illustration the vertical retrace time is assumed to be very small.) This method of operation has the consequence that for an observer at such a distance that he just fails to resolve individual scanning lines the effective flicker frequency has been doubled; thus, very much greater brightness is permissible without any increase in the picture repetition rate. This advantage is only slightly impaired when the observer's distance is such that he can commence to resolve individual scanning lines. The analogy between the practice of interlacing and the motion-picture practice of interrupting the light at a rate greater than the frame frequency should be noted.

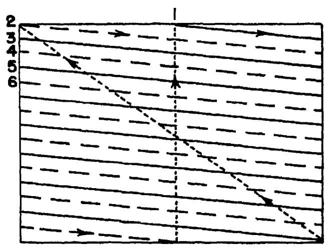


FIG. 6. Interlaced Scanning

In interlaced scanning the time required by a single vertical traverse of the picture is no longer equal to the time required to scan a complete picture. A single vertical scan is called a field; a complete picture is called a frame. The customary variety of interlace thus has two fields per frame. Higher orders of interlace have been proposed but have not found widespread use.

It is desirable that the scanning processes, both horizontal and vertical, repeat exactly from cycle to cycle. If the number of lines to a complete frame is an odd number, and a frame consists of two fields, correct interlace and uniformity of repetition of the scanning operation go hand in hand. This arrangement is called "odd-line interlace."

3. RESOLUTION AND FLICKER REQUIREMENTS; BAND WIDTH

In article 2 it was noted that the vertical width of confusion was equal to $\sqrt{2}$ times the scanning-line pitch; hence

$$W_v = \frac{V\sqrt{2}}{n} \tag{4}$$

where V is the picture height and n the number of useful scanning lines per picture. It was also noted that the horizontal width of confusion, in seconds, was

$$t = \frac{1}{2f_c} \tag{1}$$

Since

$$W_h = tv \tag{5}$$

where v is the horizontal spot velocity, it follows from eq. (1) that

$$W_h = \frac{v}{2f_c} \tag{6}$$

It is desirable that the horizontal and vertical widths of confusion be approximately equal. Thus

$$\frac{V\sqrt{2}}{n} = \frac{v}{2f_c} \tag{7}$$

whence

$$f_c = \frac{nv}{2\sqrt{2}V} \tag{8}$$

The number of useful lines in the complete picture, n, will be less than the total number of line-periods in the picture time-interval, n', by a factor a (usually 0.90 to 0.93) introduced to provide time for the vertical return of the spot. The line-repetition rate, or line-scanning frequency, in cycles per second, is the product of the picture-repetition frequency N and the total number of lines per picture, n'.

If no time were allowed for horizontal return of the scanning spot from the right to the left of the raster, the horizontal velocity would have the value

$$v = WNn' \tag{9}$$

where W is the picture width. Retrace time must be provided, thus reducing the useful portion of the line-period from unity to a fraction b, usually 0.82 to 0.85. The actual velocity is then

$$v = \frac{WNn'}{b} \tag{10}$$

Substituting this value in eq. (8)

$$f_c = \frac{nn'NW}{2\sqrt{2}bV} = \frac{an'^2N}{2\sqrt{2}b} \times \frac{W}{V} \tag{11}$$

and, since W/V, the aspect ratio, is $^4/_3$,

$$f_c = \frac{2a}{3\sqrt{2}b}\, n'^2N \tag{12}$$

If the picture is viewed from a distance equal to four times its height, the angular width of confusion is found from eq. (4) to be

$$W_a = \frac{\sqrt{2}}{4an'} \quad \text{radian} \tag{13}$$

while the corresponding square solid angle is

$$S = \frac{1}{8a^2n'^2} \quad \text{steradian} \tag{14}$$

In article 1 it was stated that the figure of confusion should not exceed 1.5×10^{-6} steradian. However, the method of defining the boundaries of the figure, in the study there referred to (reference 4), differed from that of article 2 sufficiently to require introduction of the factor 1/1.9 when this figure is applied to the preceding analysis. Using the resulting value of 0.8×10^{-6} steradian, and choosing the value 0.90 for a, eq. (14) yields the value

$$n' = 439 \text{ lines per picture}$$

The American television standards have been set at 525 lines per picture, thus more than meeting the resolution requirement in the vertical direction.

The vertical repetition rate must be at least 25 and preferably 30 per second for the complete picture in order to minimize flicker, even with interlaced scanning. It is advantageous to make the rate an integral submultiple of the power supply frequency, to eliminate disturbance of interlace by stray fields; a vertical rate of 30 per second is therefore chosen. The corresponding line frequency is 30 times 525, or 15,750 cycles per second.

The effective cutoff frequency, as defined in article 2, is obtained by substituting in eq. (12) the values $a = 0.90$, $b = 0.84$, $n' = 525$, $N = 30$. Thus

$$f_c = 0.505n'^2N = 4.18 \text{ Mc}$$

With a frequency band including components up to 4 or 4.5 Mc, the gradual cutoff required to avoid severe "overshoots" (and usually obtained automatically as a consequence of scanning spot distributions) results in an effective band width which rarely exceeds 3 Mc (reference 7). The condition of equal vertical and horizontal widths of confusion (eq. [7]) is thus not usually obtained; moderate departures from this condition are known to be of only minor importance (reference 4).

4. PICK-UP DEVICES

Early work on television employed mechanical scanning of the object or scene to be televised. The optical system focused an image on a plane at which a disk provided with apertures spaced about its periphery was interposed. A single rotation of the disk produced one complete scan of the image. The photoresponsive device located behind the disk responded instantaneously to the light transmitted by the apertures as they successively traversed the image. In a modification of this arrangement the scanning process was applied to the light which illuminated the object. This modification reduced the total amount of light incident on the subject, and the accompanying heat, very considerably.

Analysis of these mechanical scanning methods shows that the light incident upon the photoresponsive device is inversely proportional to the number of picture elements, and that, for an acceptable number of elements in the picture, it is not practicable to increase the scene lighting and the pick-up lens size sufficiently to produce a useful signal-to-noise ratio (reference 26). The limitation inherent in mechanical pick-up systems encouraged the development of photo responsive devices capable of storing light energy over the entire scanning period, thus permitting an improvement of several orders of magnitude in the signal-to-noise ratio. These devices are described in Section 15.

5. PICTURE DISPLAY DEVICES

Early picture display devices employed a light source of instantaneously controllable intensity (such as a crater-type glow-discharge lamp or a high-intensity arc whose light output was modulated by a Kerr cell) in conjunction with a mechanical scanning device similar to that described in connection with the preceding article. With arrangements of this sort each picture element is illuminated for a time interval corresponding to its own duration.

The average illumination of a picture element at the viewing screen is equal to the illumination intensity during the picture element divided by the ratio of frame duration to picture element duration. A typical value for this ratio is 300,000. In consequence of this factor, and the limited intrinsic brightnesses of convenient light sources, mechanical scanning systems as heretofore proposed have been largely supplanted by electronic scanning using cathode-ray tubes (reference 2).

In a modification of a mechanically scanned television reproducer, Scophony developed a method of storing the light-modulation information in a liquid cell in such fashion that an individual picture element could be illuminated for a period many times the element's own duration. Practical difficulties have prevented widespread use of this arrangement (references 10, 11, 12, and 13).

In a cathode-ray tube the instantaneous power concentration in the scanning spot may reach values of 10 to 1000 kw per sq in. Efficiencies of fluorescent materials in converting from electrical to luminous energy lie in the range of 5 to 10 per cent. On this basis, average highlight brightnesses of a few thousand foot-lamberts are possible. The cathode-ray tube has, therefore, become the accepted television picture-reproducing device. Cathode-ray tubes are described more fully in Section 15.

Electronic control of the opacity or alternatively the light-reflection coefficient of a surface has been employed for certain special purposes. In one form the effect of an injected electron in producing opacity in a transparent alkali halide crystal is employed. In another arrangement a layer of flakelike particles such as graphite, suspended in a fluid, is used; in the absence of electric field the particles exhibit random orientation and thus prevent light transmission, but an applied electric field makes the orientation orderly and permits light transmission in the direction of the field (references 14, 15, and 16).

SCANNING CIRCUITS FOR CATHODE-RAY DEVICES. The electron beam of the cathode-ray tube may be deflected by either electric or magnetic field. Both fields involve the storage of energy in the deflection space within the tube and incidentally in the external circuits.

Wave forms for a deflecting field are shown in Fig. 7. Curve *A* shows an ideal form. Curve *B* shows a departure from the ideal form introduced to permit a finite retrace time and thus avoid the necessity for handling unreasonably excessive currents or voltages in the deflection circuits during the retrace intervals. Curve *C* shows a typical practical curve in which the form of variation is modified from that of curve *B* to eliminate the discontinuities of slope exhibited by curve *B* and thus reduce the number of harmonics which must be faithfully transmitted to the deflecting circuit.

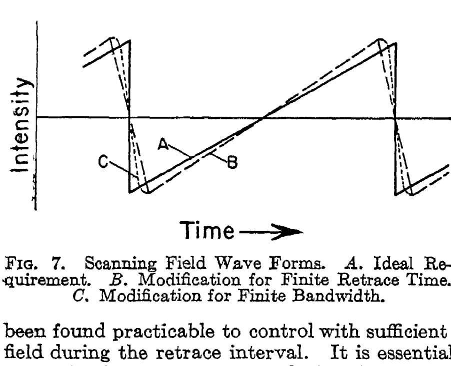


FIG. 7. Scanning Field Wave Forms. *A*. Ideal Requirement. *B*. Modification for Finite Retrace Time. *C*. Modification for Finite Bandwidth.

For faithful reproduction it is essential that the scanning wave have the same form in both the pick-up device and the reproducing device. It has not been found practicable to control with sufficient accuracy the wave form of the deflecting field during the retrace interval. It is essential, therefore, that no attempt be made to transmit picture information during this interval. Means are customarily provided for preventing the appearance of the reproducing screen from being affected by signal potentials during the retrace.

SCANNING CIRCUITS FOR ELECTROSTATIC DEFLECTION. Figure 8 shows a typical circuit for producing electrostatic deflection voltages. Oscillator tube *V*-1 is here shown as a blocking oscillator. Other forms such as multivibrators and thyratrons may also be used. The oscillator acts periodically to discharge capacitor C_1. The capacitor is then charged through R_1 until the voltage at the oscillator anode rises to a point where, in conjunction with any voltage which may be applied on the grid, oscillation is again produced, thus again discharging C_1. The voltage variation with time on C_1 during the trace portion of the period is, of course, exponential. It is essential, therefore, that the amplitude at this point be kept small enough to preserve a good approximation to

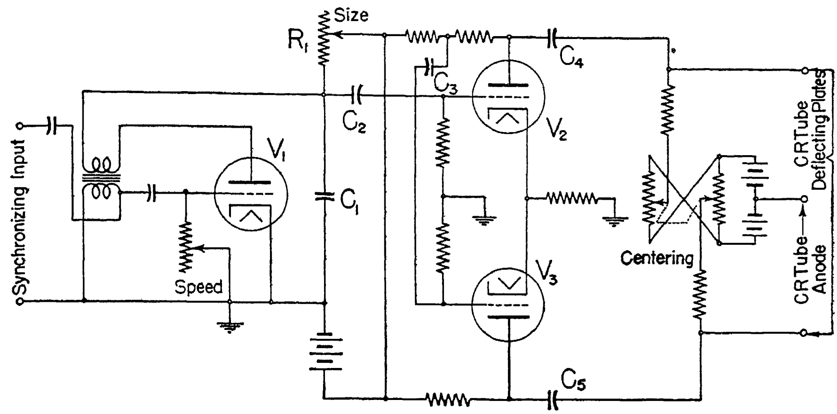


FIG. 8. Electrostatic Deflection Circuit

linearity. The amplifier tube V_2 and phase inverter tube V_3 amplify the voltage appearing across C_1 to the required level and provide the usually necessary outputs of opposite polarities. In choosing the values of components care must be taken that the capacitors C_2, C_3, C_4, and C_5 do not introduce excessive phase shift for the fundamental frequency component. A total shift of 3° is a useful upper limit. By care in the choice of component values the circuit may be made to work down to frequencies much lower than the normal television field frequency of 60 cycles; it may also be used readily at frequencies at least ten times higher than the normal line-scanning frequency of 15.7 kc.

A simpler circuit which is useful for scanning frequencies of the same order as television line frequencies is shown in Fig. 9. In this circuit capacitors C_1 and C_2 are charged from

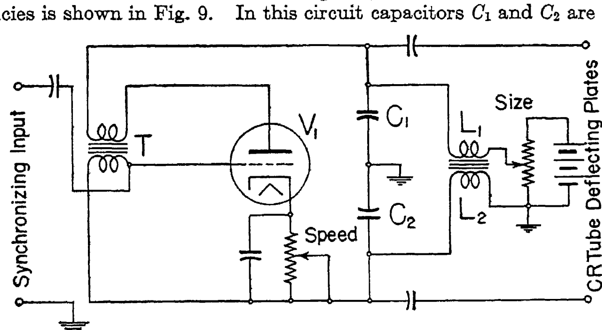


FIG. 9. Electrostatic Deflection Circuit

the power supply during the trace period through the reactor L_1, L_2. While this charging current is oscillatory, the choice of a resonant frequency less than one-tenth the scanning frequency for the circuit consisting of the two reactor windings and two capacitors in series results in the use of only the linear central portion of the oscillatory cycle. When the voltage appearing across C_1 and C_2 is great enough, tube V_1 goes into oscillation drawing a heavy current and discharging the pair of capacitors. If the equivalent series resistance of V_1 is sufficiently low, this discharge will be oscillatory and will last for one-half cycle at a frequency determined by the apparent inductance of the primary of transformer T and the capacitance of capacitors C_1 and C_2. (Insufficient leakage inductance in T makes the retrace time unnecessarily short and the peak current and dissipation in

V_1 excessive.) The potential difference between the opposite ends of C_1 and C_2 reverses during this discharge and rises to a negative value which may be a considerable fraction of the voltage on the capacitors before V_1 began to conduct. With the cessation of conduction in V_1, C_1 and C_2 are again charged through L_1, L_2, and the cycle repeats.

CIRCUITS FOR MAGNETIC SCANNING. In the magnetic scanning cycle the current in the coil by which the magnetic field is produced is zero at the midpoint of the scanning trace. During the latter half of the trace the current is built up under control of a vacuum tube to the value required to produce the full deflection. The flow of current through the vacuum tube is then interrupted, and the scanning current continues to flow in the coil, charging the distributed capacitance which is in shunt with the coil until the current vanishes. The distributed capacitance is now charged to a high potential; it discharges back into the coil, producing a current of opposite polarity to that which previously flowed. When the potential energy stored in the distributed capacitance has been transferred completely back to the coil, the next scanning-line trace starts; the flow of coil current is permitted to continue through a vacuum tube in order to cause the current to decrease linearly to zero value at the midpoint of the trace. In this cycle energy is supplied to the circuit during the latter half of each trace, a half-period of free oscillation takes place during the retrace, and the energy is then dissipated during the first half of the succeeding trace. These relations are shown on an ideal basis in Fig. 10. It is necessary, of course, that the resonant half-period of the free oscillation not exceed the permissible retrace time.

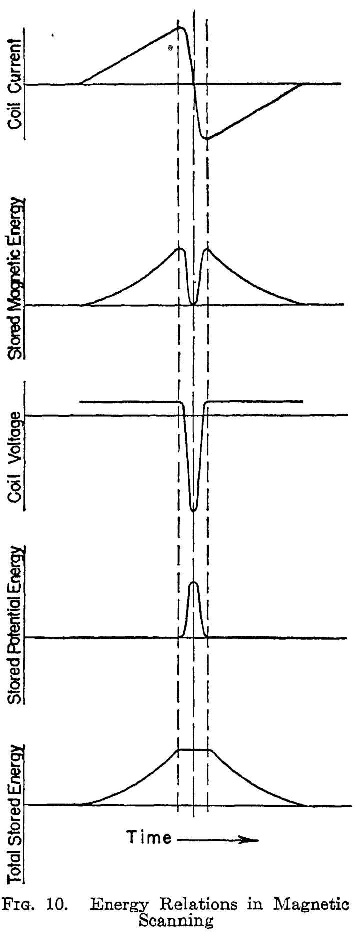


FIG. 10. Energy Relations in Magnetic Scanning

The energy required in the magnetic field (within the tube) to deflect a cathode-ray tube fully is a function of accelerating voltage and deflection angle primarily. Typical tubes require amounts of energy between 30 and 300 microjoules. These amounts of energy must be provided and dissipated during each scanning cycle. Thus, for a type 10BP4 cathode-ray tube requiring about 300 microjoules for horizontal deflection and a horizontal deflection frequency of 15,750 the power which must be delivered by the scanning circuit within the tube is 4.8 watts.

A typical line-frequency magnetic scanning circuit is shown in Fig. 11.

A step-down transformer is interposed between the energy supply tube V_1 and the scanning coil or yoke Y, to decrease the effective distributed capacitance of the circuit, thus assisting to obtain suitably short retrace time. The wave-form diagrams of the figure, which are corrected for the effect of the transformer ratio, show the current gradually increasing in tube V_1 during the trace and terminating abruptly at retrace. During this same period the current in tube V_2 rises abruptly to a maximum value and decreases smoothly during the rest of the trace. The analogy between these two current wave forms and the corresponding forms in a class AB amplifier should be noted. Tube V_3 provides the feedback connection for self-oscillation.

The circuit is arranged to make use of the energy returned by the yoke Y during the first half of the trace. This energy is stored in capacitor C_1; the voltage thus developed across C_1 is used to augment the anode supply voltage for tube V_1. This is called the "bootstrap" connection.

It is usually necessary, for line-frequency scanning, to use a relatively low inductance in the yoke. The current is therefore relatively high and may reach peak values of 0.1 to 1.0 amp. Relatively large direct currents (0.01 to 0.20 amp) are consequently required to correct any incidental decentering of the raster. The circuit of Fig. 11 is therefore designed to permit the full B supply current of the receiver to flow through the centering control.

Overall efficiency of the circuit of Fig. 11 may reach a value of 15 per cent. Delivery of 5 watts to the magnetic field within the tube thus requires at least 30 watts of d-c power (reference 80).

The power required for the relatively slow vertical scanning is very much smaller than that for horizontal scanning. A typical circuit is described in article 32 of this section.

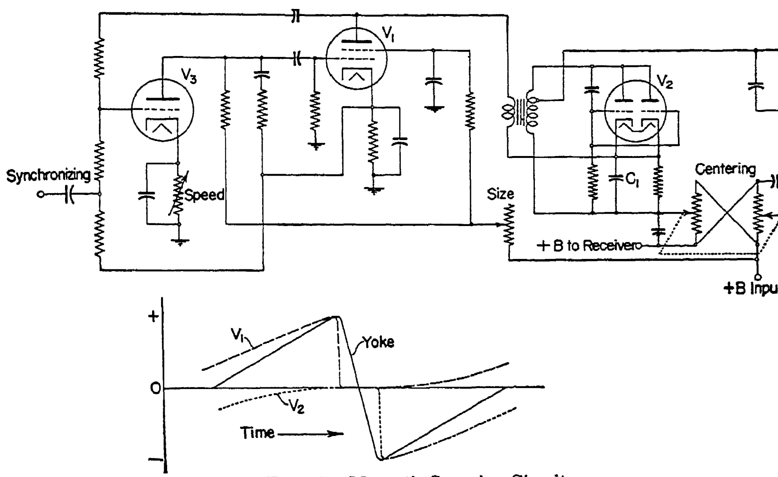


FIG. 11. Magnetic Scanning Circuit

KEYSTONE CORRECTION. When the electron beam in its undeflected position does not strike its target at a right angle the shape of the raster produced by uniform deflecting fields will be a keystone. Circuit arrangements which modulate one of the scanning generators by a signal derived from the other are required to produce a rectangular raster under these conditions. Application for these is found in the Iconoscope type of camera tube and in certain forms of picture tubes (reference 81).

6. SYNCHRONIZING

In television practice the picture information is generated in an orderly sequence. The picture display device must display this information in the same sequence if the original picture is to be reproduced. It is necessary, therefore, that information to synchronize the scanning operations of the display device be furnished with the picture information and that this information be subject to delays in transmission identical to those experienced by the picture information. Synchronizing signals are, therefore, included with the picture signals.

There are two ways in which scanning devices may be synchronized. In the simpler of these the synchronizing signal has essentially a pulse form and is applied to the scanning device in such fashion as to terminate the scanning trace and initiate the retrace. This action takes place at a speed limited only by the transient response of the scanning oscillator itself. This method of operation has the advantage of simplicity but the disadvantage that the scanning cycle may be mistimed and the picture consequently distorted by either: (1) a noise impulse tripping the oscillator prematurely; (2) loss of a synchronizing pulse due to a temporary blocking of the signal channel by noise, or (3) the combination of random noise components with the synchronizing pulse to produce random phase variation of the leading edge of the pulse.

The alternative synchronizing method is to apply the synchronizing signal and a signal derived from the scanning device to a phase comparison circuit whose output voltage controls the frequency of the scanning device. If the synchronizing pulses are uniformly spaced and the scanning device is itself stable in frequency, the phase control may be made slow acting, thus effectively decreasing the band width of the synchronizing signal channel and reducing its susceptibility to noise interference by several orders of magnitude (reference 17).

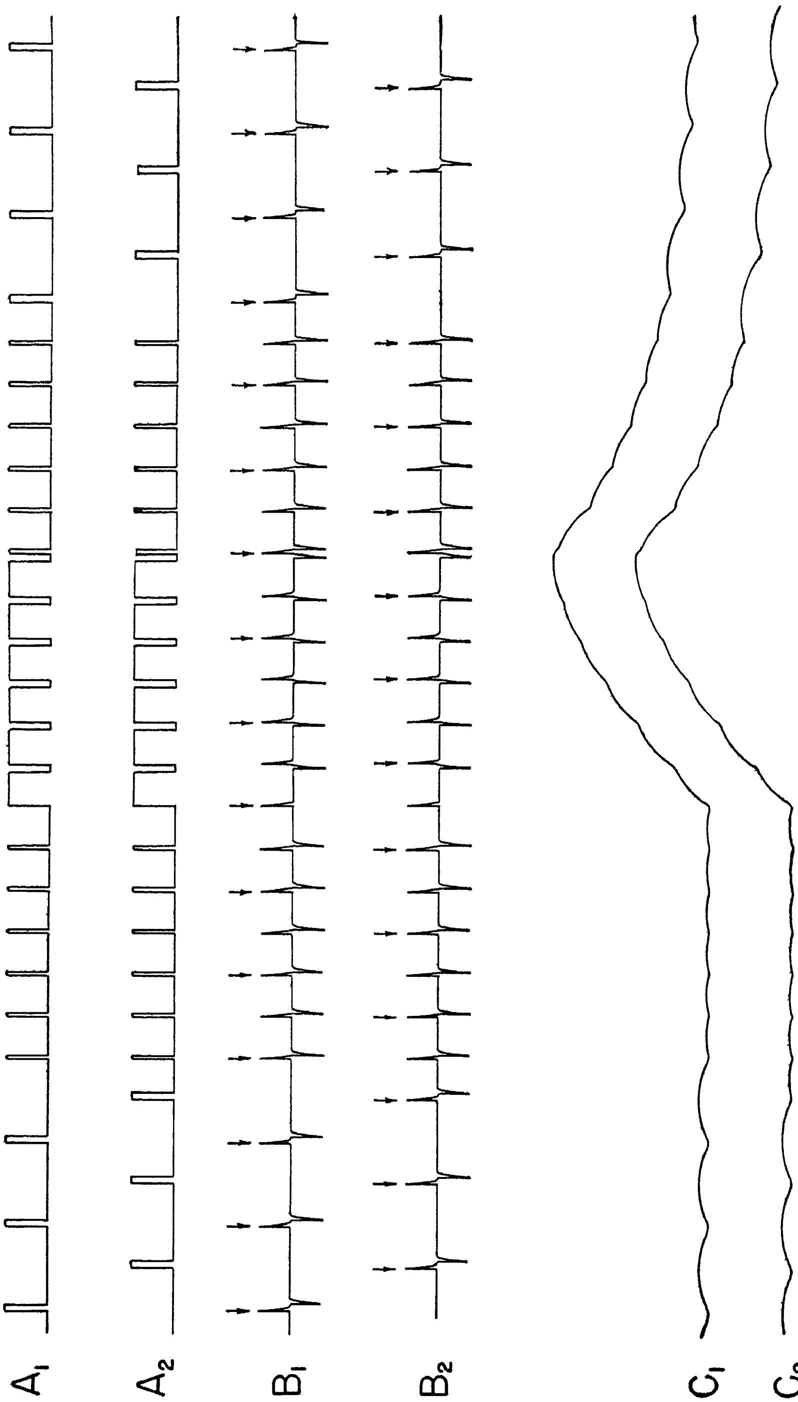
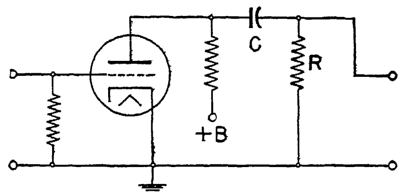
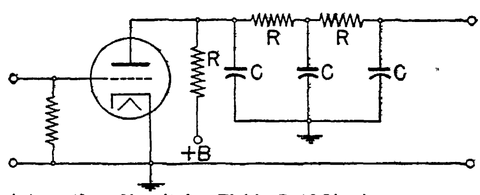


FIG. 12. Synchronizing Signal Wave Forms. *A*. Composite. *B*. Effect of Differentiating. *C*. Effect of Integrating.

Separate synchronizing signals are required for the two directions of scanning. With interlaced scanning it is essential that these signals be readily separable one from the other in a receiver. The form of composite synchronizing signal which has been adopted to meet these requirements is shown in Fig. 12. In line A_1 of Fig. 12 the last three line-synchronizing pulses preceding the vertical retrace interval appear at the left, followed by six equalizing pulses, six field synchronizing or "broad" pulses, and six more equalizing pulses. Transmission of normal line pulses is then resumed, to continue until the next field retrace interval, shown at A_2. The traces A_1 and A_2 are separated in time by *exactly* one field period; in consequence of the use of "odd-line interlace," the line pulses in A_1 and A_2 are one-half line period out of phase. The line-frequency signals may be separated from the composite signal by differentiation, as shown at B_1, B_2; the arrows here represent the times at which the line oscillator should "fire"; at each of these, a synchronizing signal is provided. The field frequency signals may be separated by integration, with the results shown at C_1, C_2. It is a characteristic of an integrating circuit that it "remembers." For this reason the interval immediately preceding the vertical synchronizing signal contains horizontal synchronizing signals at twice the normal repetition rate. By this means the time intervals immediately preceding the vertical synchronizing pulses in the two fields are made identical. The line synchronizing pulses are reduced to half their normal duration during this period so that their integrated value will be no greater than that of line pulses of normal duration and normal repetition rate. These equalizing pulses also appear for a short interval following the vertical synchronizing signal to insure that during the entire interval in which the vertical scanning device is sensitive to synchronizing signals those signals will be alike in both fields.



Differentiating Circuit for Line; RC = 2 μsec.



Integrating Circuit for Field; R=100k ohms, C=0.005 μf.

FIG. 13. Synchronizing Signal Selection

The diagrams of Fig. 13 illustrate differentiating and integrating circuits to perform the separations shown in Fig. 12.

7. THE VIDEO SIGNAL

The video signal is generated by a pick-up tube as described in Section 15. The output of this device usually requires amplification to raise it to usable level and may, in addition, require processing to remove from the signal certain spurious components which are not properly a part of the signal. Its direct component must either be transmitted faithfully with the same gain as other components or be reinserted by either manual or automatic means after amplification has taken place.

Figure 14 shows a test pattern used for testing television systems and the oscillogram of a single scanning line of that pattern. The oscillogram was taken at a point in the transmission system where all spurious components had been eliminated. The correspondence between elements of the picture and elements of the oscillogram is shown.

TRANSMISSION OF THE D-C COMPONENT. It is theoretically possible to transmit and amplify the d-c component along with the other components of the video signal. In practice, however, this is frequently found to be inconvenient. A satisfactory alternative, known as "d-c reinsertion," may be followed once the black level of the signal has been established. In this alternative practice the recurring black intervals are used to provide an a-c carrier of the d-c component. This practice is illustrated in its simplest form in the two-stage amplifier of Fig. 15. The blocking capacitor C_1 prevents transmission of the d-c component from the anode circuit of V_1 to the grid of V_2. The video signal is applied to the grid of V_1 with positive polarity, that is, with an increase in object brightness represented by a change of signal potential in the positive direction. The black level, therefore, represents the most negative portion of the signal. The signal polarity is reversed in the anode circuit of V_1 so that at this point the black level is the most positive portion of the signal. The diode V_3 is connected to conduct on the positive portion of

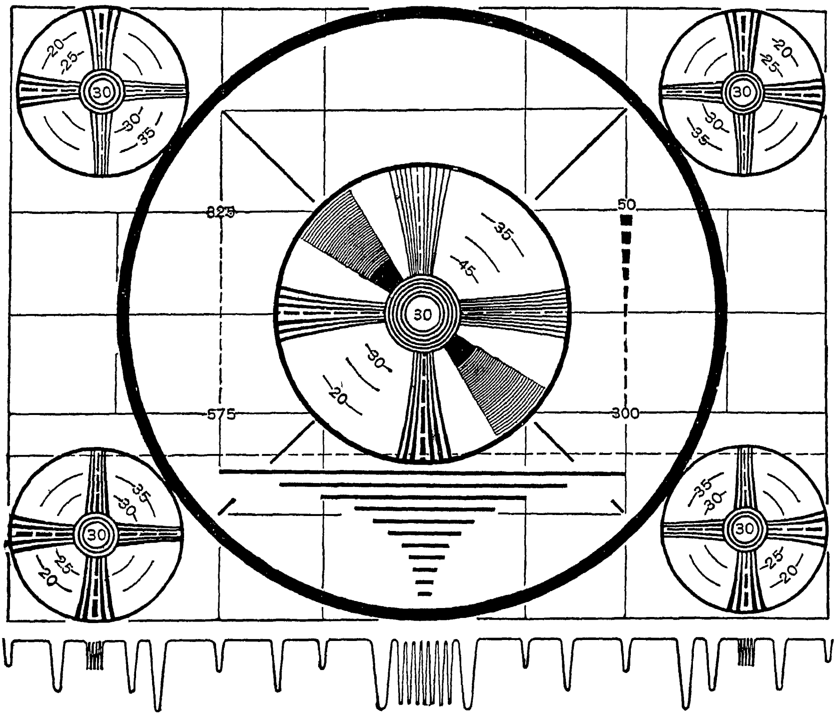


Fig. 14. Television Test Pattern and Sample Line Wave Form. Test pattern (copyrighted by Radio Corp. of America). Wave form of single line as shown.

the signal reaching it. (The grid-to-cathode conductance of V_2 shows a similar characteristic.) Current flowing through the diode in response to the black intervals of the signal establishes such a charge on C_1 as to reduce the flow of current through the diode to a value just sufficient to make up for the leakage from the condenser through R_1. If the effective resistance of the charging path (including the series resistance of the diode and the resistance of R_2) is small, the amount by which the diode anode is positive with respect to its

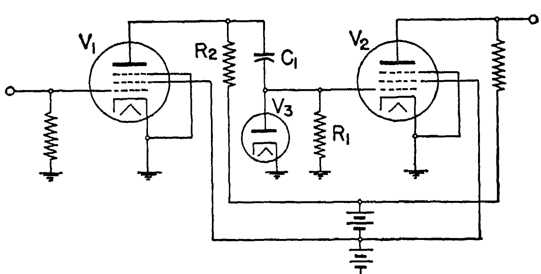


Fig. 15. D-c Reinsertion Circuit

cathode, during the black intervals, will be a negligible fraction of the total signal potential. The black intervals of the signal will, therefore, be held at substantially the potential of the diode cathode, and the gray and white portions will extend negatively from this potential in their appropriate amounts.

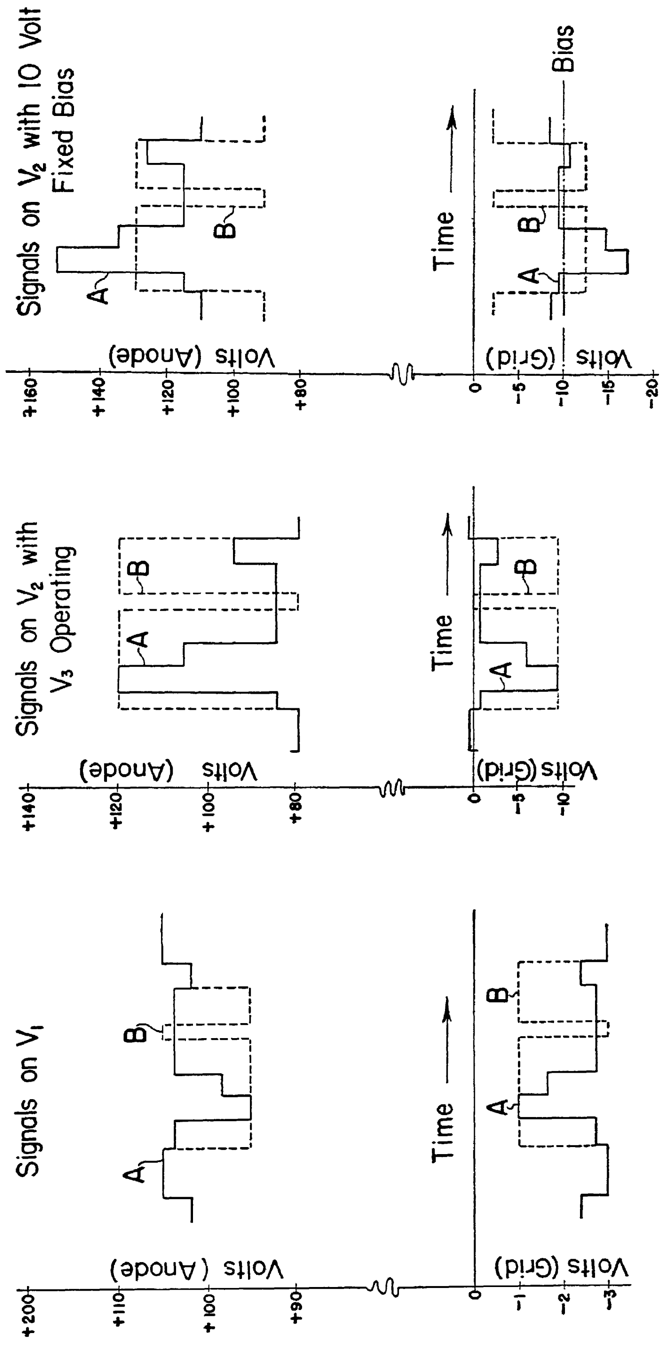


FIG. 16. D-c Reinsertion

Figure 16 shows the performance of the circuit of Fig. 15 with two different signals. Signal A has a little white, but mostly dark gray and black, while signal B is all white except for a brief black interval. The effectiveness of the reinsertion action is shown by the substantial agreement in black levels for both signals, at the grid and also at the anode of V_2. The curves shown for the performance at V_2 when the diode is replaced by a fixed bias show the much greater range of voltages which tube V_2 must handle without distortion, if d-c reinsertion is not practiced.

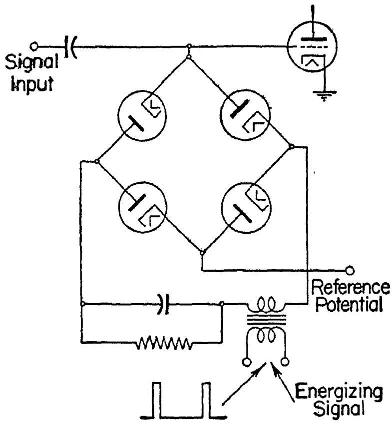


FIG. 17. Clamp Circuit

Because of the high impedance of diodes at the low currents they are called upon to handle in the circuit of Fig. 15, the accuracy with which the black level is maintained is somewhat imperfect. When more accurate performance is required, a more elaborate arrangement known as a clamp circuit may be used in place of the diode V_3 and resistor R_1 of Fig. 15. In a clamp circuit, an external source applies power to a network of diodes, producing considerable currents and low diode impedances. The network is balanced so that any deviation of the signal potential from the potential of a reference point during the flow of current from the external source is corrected by a small unbalance in the currents in the branches of the circuit. Figure 17 shows one form of clamp circuit. Clamp circuits differ from the circuit Fig. 15 in that (*a*) they are capable of conduction in both directions; (*b*) in consequence of this they must be energized only during appropriate time intervals. See also reference 27 and 84.

8. THE COMPOSITE SIGNAL

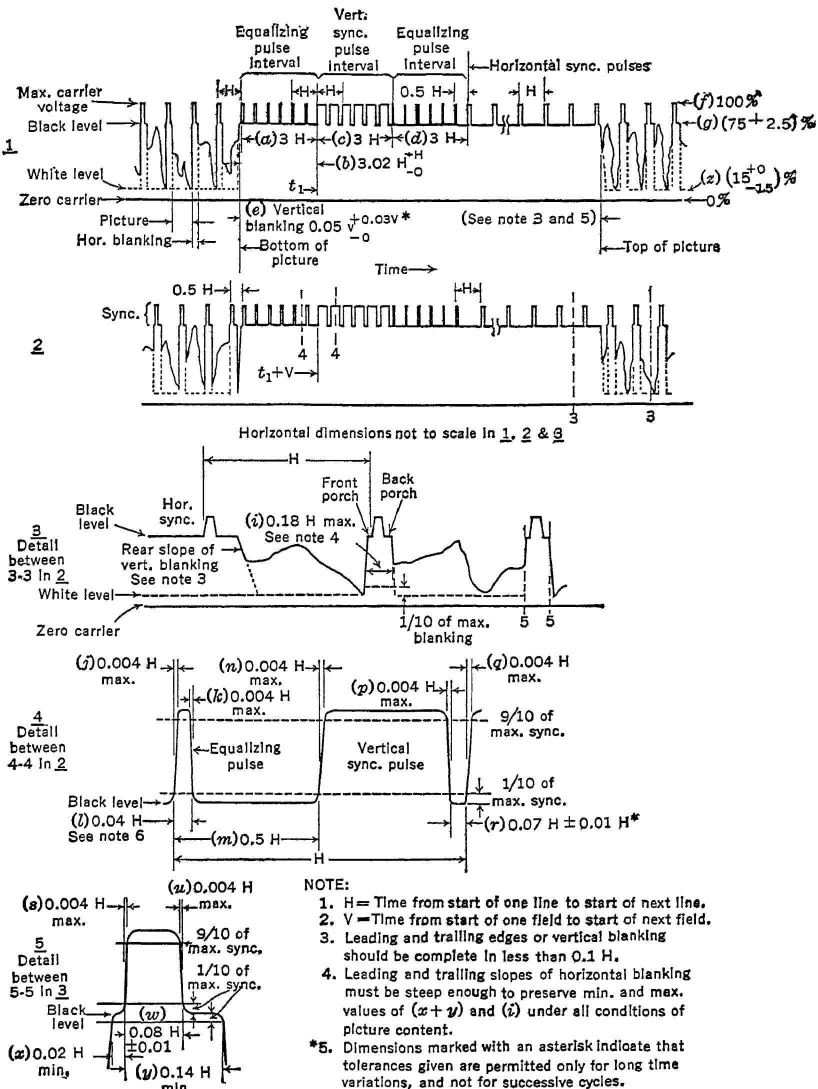
In the design of television systems provision must be made for the transmission of these four signals: (*a*) video signal; (*b*) horizontal synchronizing signal; (*c*) vertical synchronizing signal; and (*d*) sound signal.

The system may be designed to transmit all four of these from separate transmitters. Alternatively, two or more may be combined and transmitted by a single transmitter. The combination of the picture signal and the two synchronizing signals in a single transmission has been recognized as particularly suitable, since it simplifies both receiving and transmitting apparatus and also removes some sources of non-uniform transmission delay between these components.

The construction of a composite signal containing these three individual signals requires the synchronizing and picture signals to occupy different ranges of amplitude, since these two classes of signals cannot be distinguished from one another by a frequency separation. They must also occupy different time intervals. These requirements are satisfied by assigning a range of potentials beyond black (and, therefore, called infra-black) to the synchronizing signals and by inserting synchronizing signals in the time intervals provided for scanning retrace. Figure 18 shows at 3 an oscillogram of two lines of a composite signal showing line synchronizing pulses properly located in the retrace intervals. The position of the leading edge of the pulse in the retrace interval is set a short time after the beginning of the interval so that even receiver circuits of somewhat restricted band width (and hence slow transient response) will have time to reach black level before the synchronizing pulse begins, regardless of whether the picture edge is white or black. Failure to provide this interval results in phase modulation of the scanning by the picture content. This interval (sometimes called the "front porch") is made no greater than required by the foregoing consideration, since any extra waiting at this point is at the expense of either decreased time available for scanning retrace or decreased time available for picture.

The placement of the field signal conforms to the practice already described for the line signal. The field signal is located in the same region of amplitude as the line signal and is placed to occur during the vertical retrace region.

The portion of the transmission amplitude range not occupied by synchronizing signals is reserved for the picture information.



NOTE:

1. H = Time from start of one line to start of next line.
2. V = Time from start of one field to start of next field.
3. Leading and trailing edges of vertical blanking should be complete in less than 0.1 H.
4. Leading and trailing slopes of horizontal blanking must be steep enough to preserve min. and max. values of $(x + y)$ and (i) under all conditions of picture content.
*5. Dimensions marked with an asterisk indicate that tolerances given are permitted only for long time variations, and not for successive cycles.
6. Equalizing pulse area shall be between 0.45 and 0.5 of the area of a horizontal sync. pulse.
7. Refer to text for further explanations and tolerances.

FIG. 18. FCC Standard Television Synchronizing Wave Form

9. THE RADIO-FREQUENCY SIGNAL

The composite signal of Fig. 18 may be applied to an r-f carrier as either amplitude, phase, or frequency modulation. In television broadcasting, multipath transmission is frequently observed; picture distortions caused by multipath transmission when phase or frequency modulation is used are so serious that these methods of modulation have not seemed practicable. Television broadcasting, therefore, makes use of amplitude modulation.

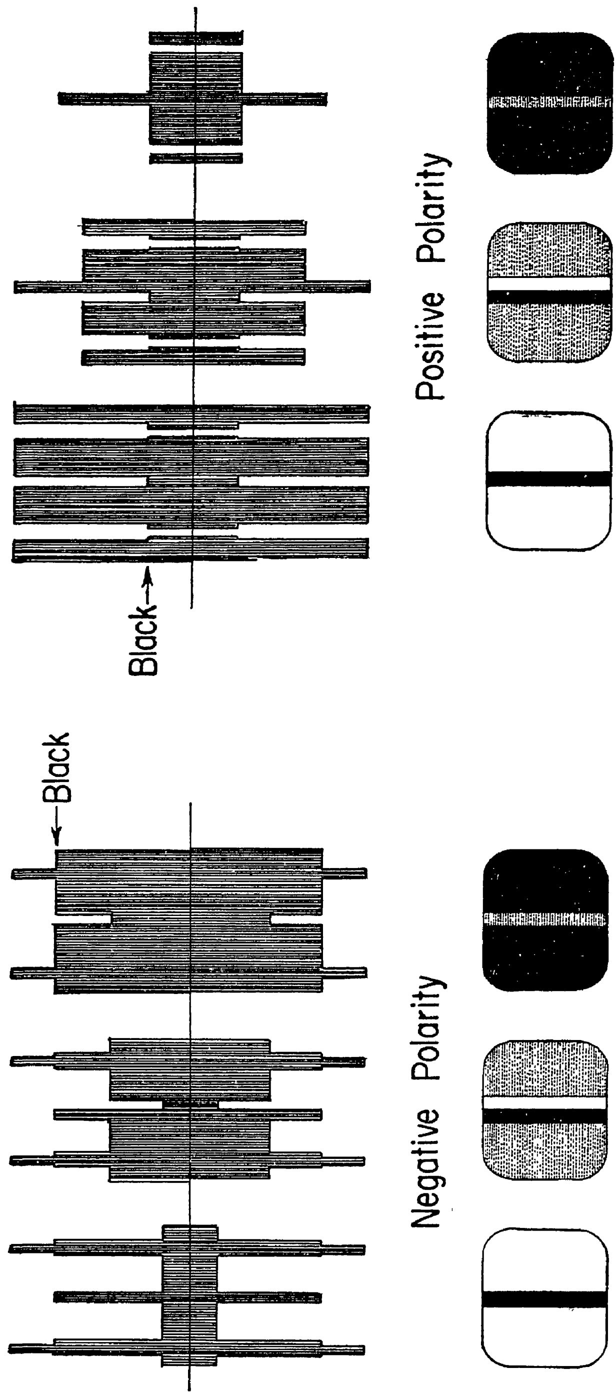


Fig. 19. Modulation Envelope Wave Forms

POLARITY OF MODULATION. Polarity of modulation may be either positive (that is, with an increase of image brightness represented by an increase of radiated signal) or negative. Figure 19 illustrates these two forms, showing individual scanning line signals for three distributions of picture content. A positive modulation polarity signal includes at all times the synchronizing level (zero carrier) and the black level. It does not indicate the level of peak white unless elements of this intensity are present in the picture. Negative modulation polarity, on the other hand, includes at all times the synchronizing level (maximum carrier intensity), the black level, and peak white (zero carrier).

Automatic gain control circuits for receivers require the presence in the received signal of some characteristic which is independent of modulation. In sound transmissions, the average value of the carrier has the required characteristic, but in television signals, the average value is dependent on average picture brightness. White level, black level, or synchronizing level must be used instead. Preferably, the peaks of the signal envelope should be used, so that a simple peak detector may serve as the source of automatic gain control information. It is found, therefore, that negative modulation polarity simplifies very much the provision of automatic gain control in receivers.

The effects of impulse noise interference on signals of the two polarities are quite different. With positive modulation impulse noise usually produces bright spots in the reproduced picture and has little effect on synchronizing signals. With negative modulation impulse noise produces primarily black spots on the picture (which are on the whole less disturbing than the bright spots produced with positive modulation) but has a greater tendency to interfere with synchronizing signals. Since it is found possible to minimize the effect of impulse noise on synchronizing sufficiently by careful circuit design in the receiver and since automatic gain control is believed desirable, American standards for television have chosen negative modulation polarity.

BAND WIDTH. As was shown in article 3, the desired band width of television video signals exceeds 4 megacycles. The application of this signal as amplitude modulation to a carrier produces a signal having a total spectrum width which exceeds 8 megacycles. Since radio channels are not available in unlimited quantities and since also the cost of amplifiers is increased as their required band width increases, television broadcast practice is based on vestigial sideband transmission. Curves *A* and *B* of Fig. 20 show the radio-

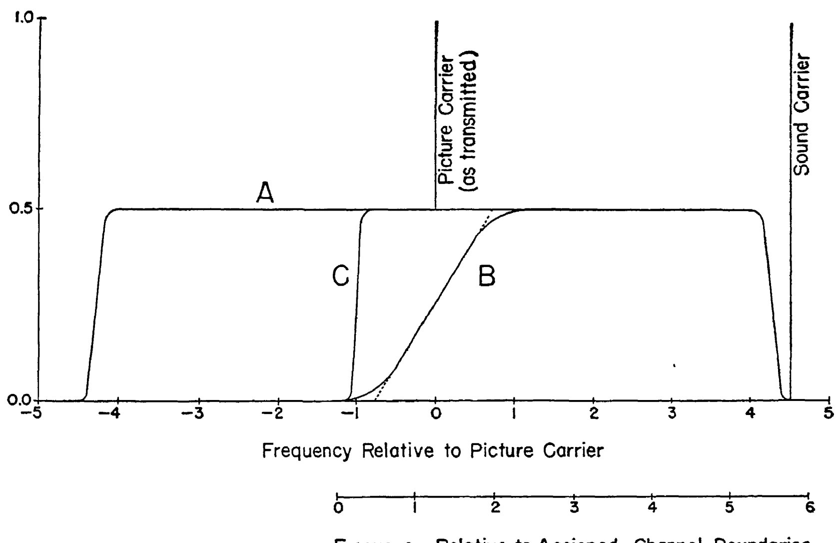


FIG. 20. Radio-frequency Amplitude Characteristics. *A*. Double Side-band Transmission. *B*. Vestigial Side-band Transmission Overall Characteristic. *C*. Vestigial Side-band Transmission; Transmitter Only.

frequency amplitude characteristics for double sideband transmission and for vestigial sideband transmission (references 19 to 24).

The overall transmission characteristic for vestigial sideband transmission requires that the signal originally produced with a carrier and symmetrical sidebands must be attenuated selectively. The practice which has been standardized is to provide a receiver characteristic having the same form as the desired system overall characteristic. The corresponding transmitter characteristic must, therefore, exhibit negligible attenuation at all frequencies which are effectively transmitted by the receiver. It has therefore been standardized as shown in Fig. 20 as curve *C*.

SOUND TRANSMISSION. The sound accompanying a television picture is transmitted on a separate carrier whose frequency is located, with respect to the picture carrier and its sidebands, as shown in Fig. 20. The sound carrier is frequency modulated with maximum deviation of 25 kilocycles. The pre-emphasis practice standardized for frequency-modulated sound broadcasting is also used for television sound.

FREQUENCY ALLOCATION. Because of the wide frequency channels required for television, allocation of channels below 50 megacycles would interfere with so many existing services as to be impracticable. Television allocations for commercial use, therefore, lie in the range between 54 and 216 megacycles, as shown in the table. Allocations at higher frequencies have been made for experimental and relay use.

Channel	Nominal Boundaries	Picture Carrier	Sound Carrier
2	54– 60	55.25	59.75
3	60– 66	61.25	65.75
4	66– 72	67.25	71.75
5	76– 82	77.25	81.75
6	82– 88	83.25	87.75
7	174–180	175.25	179.75
8	180–186	181.25	185.75
9	186–192	187.25	191.75
10	192–198	193.25	197.75
11	198–204	199.25	203.75
12	204–210	205.25	209.75
13	210–216	211.25	215.75

POLARIZATION OF RADIATED SIGNAL. A simple horizontally polarized dipole antenna has a horizontal directive pattern which is sometimes useful in minimizing effects of multiple transmission paths. Since in other respects there is little net advantage either way between horizontal and vertical polarization, horizontal polarization has been standardized.

10. STANDARDS

The Federal Communications Commission has established the following Standards of Good Engineering Practice for television broadcasting:

1. The width of the television broadcast channel shall be 6 megacycles per second.
2. The visual carrier shall be located 4.5 megacycles lower in frequency than the aural center frequency.
3. The aural center frequency shall be located 0.25 megacycle lower than the upper frequency limit of the channel.
4. The visual transmission amplitude characteristic shall be as shown in Appendix II [curve *C* of Fig. 20].
5. The number of scanning lines per frame period shall be 525, interlaced 2:1.
6. The frame frequency shall be 30 per second, and the field frequency shall be 60 per second.
7. The aspect ratio of the transmitted television picture shall be 4 units horizontally to 3 units vertically.
8. During active scanning intervals, the scene shall be scanned from left to right horizontally and from top to bottom vertically, at uniform velocities.
9. A carrier shall be modulated within a single television channel for both picture and synchronizing signals, the two signals comprising different modulation ranges in amplitude (see Appendices I and II) [Figs. 18 and 20].
10. A decrease in initial light intensity shall cause an increase in radiated power (negative transmission).
11. The black level shall be represented by a definite carrier level, independent of light and shade in the picture.
12. The pedestal level (normal black level) shall be transmitted at 75 per cent (with a tolerance of plus or minus 2.5 per cent) of the peak carrier amplitude.
13. The maximum white level shall be 15 per cent or less of the peak carrier amplitude.
14. The signals radiated shall have horizontal polarization.
15. A radiated power of the aural transmitter not less than 50 per cent or more than 150 per cent of the peak radiated power of the video transmitter shall be employed.

16.* *Variation of Output.* The peak-to-peak variation of transmitter output within one frame of video signal due to all causes, including hum, noise, and low-frequency response, measured at both synchronizing peak and pedestal level, shall not exceed 5 per cent of the average synchronizing peak signal amplitude.

17.* *Black Level.* The black level should be made as nearly equal to the pedestal level as the state of the art will permit. If they are made essentially equal, satisfactory operation will result and improved techniques will later lead to the establishment of the tolerance if necessary.

18.* *Brightness Characteristics.* The transmitter output shall vary in substantially inverse logarithmic relation to the brightness of the subject. No tolerances are set at this time.

See also reference 28 for an extended discussion of standards.

TELEVISION BROADCASTING

By T. J. Buzalski, A. L. Hammerschmidt, and F. J. Somers

A modern television broadcasting plant provides facilities for the pick-up and broadcast of entertainment, news, and cultural and educational subject matter in both sight and sound. The purpose of such a plant is to provide an adequate and satisfactory public service, and this requires a flexible and well-coordinated installation. A functional subdivision of equipment and facilities is the following:

(*a*) Studio and control facilities (picture and sound).
(*b*) Field pick-up and relay facilities (picture and sound).
(*c*) Visual and aural broadcast transmitters.

Typical studio and control facilities consist of one or more live-talent studios, a film pick-up studio, a video effects studio, one or more announcers' booths, and a master control room having switching and monitoring facilities for feeding the various studio outputs or remote pick-up outputs to the transmitter, as required. The master timing or synchronizing generator, various picture line amplifiers, power supply rectifiers, and other equipment common to the studio facilities system are usually grouped in a main equipment room for maximum efficiency and ease of maintenance.

Field pick-up facilities include portable television cameras with their associated control, monitoring, and synchronizing equipment and portable sound equipment. Either radio relay circuits, coaxial cables, or equalized telephone lines are used to transmit the picture signals back to the master control room of the broadcast station proper. The sound portion of the program is generally transmitted back by wire line, though radio circuits are used where wire facilities are not available. Field pick-ups also encompass the use of mobile equipment where the television cameras, along with their synchronizing, control, and monitoring equipment, are mounted in a moving boat, aircraft, or other means of locomotion.

The need for maximum height of the transmitting antenna to provide line-of-sight reception for as many receivers as possible usually requires that the visual and aural transmitters be located remote from the television studios. The visual link between the master control switching point and the transmitter may be a radio relay circuit, a coaxial cable, or an equalized telephone line. The aural link between the master control point and the transmitter is usually a wire line.

In order to coordinate operations and to assure program continuity, the television plant must be provided with an adequate and flexible intercommunication and order wire system separate and apart from the sound program pick-up, control, and transmission equipment.

11. LENS APERTURE REQUIRED

The lens speed required for a given camera pick-up tube and scene illumination is best determined by experiment under operating conditions or from data supplied by the tube manufacturer. If it is desired to compute the lens speed which will provide a sufficiently bright image to meet the requirements of a given pick-up tube, the following formulas (see reference 33) are applicable:

$$F \text{ (focal length of lens)} = \frac{W}{2} \cot \frac{\alpha}{2} \text{ inches} \tag{1}$$

$$f \text{ (numerical aperture)} = \frac{0.064W\sqrt{TBs}}{\sqrt{10^6 I_n N}} \tag{2}$$

* These items are subject to change but are considered the best practice under the present state of the art. They will not be enforced pending a further determination thereof.

where W = the width in inches of the pick-up tube sensitive surface.
α = the desired horizontal angle of view in degrees.
s = the pick-up tube sensitivity in signal microamperes per lumen.
B = the surface brightness of the scene in candles per square foot.
T = the light transmission factor of the lens (usually between 40 and 60 per cent).
N = the required peak picture signal-to-rms noise ratio.
$\bar{I}_n$ = the equivalent rms noise current (amperes) at the input of the amplifier used with the pick-up tube.

The noise generated in pick-up tubes without electron multipliers is small compared with that originating in the first tube of the video-preamplifier. Noise currents generated in the first video amplifier stage result from two significant components: thermal agitation noise in the input circuit and current fluctuations in the plate circuit of the first video amplifier tube. An expression combining these components for the computation of the equivalent rms noise current ($\bar{I}_n$) follows (see reference 33):

$$\bar{I}_n = 2\sqrt{\frac{kTfm}{R}\left(1 + \frac{RR_t(2\pi fmC)^2}{3}\right)} \tag{3}$$

where $\bar{I}_n$ = equivalent rms noise current in amperes.
k = Boltzmann constant (1.37×10^{-23} joule per °K).
T = absolute temperature (300° K).
fm = pass band of amplifier in cycles per second.
R = input resistor of amplifier in ohms.
R_t = equivalent grid resistance of input tube for noise in ohms at 300° K (for first video amplifier tube).
C = total shunt capacity (pick-up tube, stray capacity, and video amplifier input capacity).

In a non-storage pick-up tube where the photoelectrons are amplified by an electron multiplier incorporated in the pick-up tube, the predominant noise originates at the photocathode. (See reference 34.) The equivalent rms noise current in this case is given by the following expression:

$$\bar{I}_n = \sqrt{2ei_0fm} \text{ amperes} \tag{4}$$

where i_0 = current for one picture element.
e = electron charge = 1.59×10^{-19} coulomb.
fm = pass band of equipment (cycles per second).

It should be pointed out that the sensitivity s of a pick-up tube in signal microamperes per lumen is not necessarily a constant. In some types of tubes s is a function of the incident light. This is illustrated by the curve of Fig. 1, which shows s for a sample Icono-

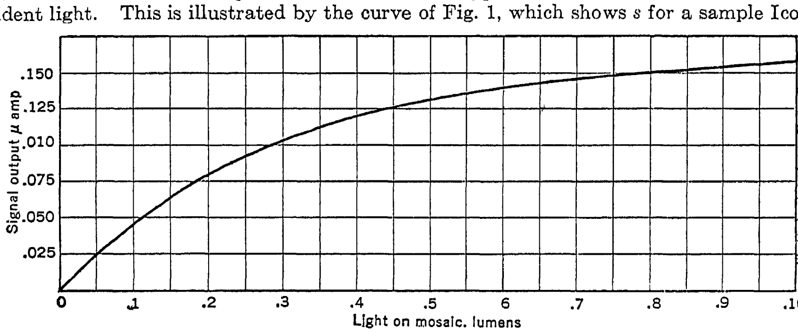


FIG. 1. Signal Output vs Illumination Characteristic of a Typical Iconoscope

scope (reference 56). Another type of tube having a non-linear sensitivity curve is the Image Orthicon (reference 30). The signal output vs. illumination curves of the Orthicon and the Image Dissector are essentially straight lines.

12. STUDIO CAMERA DESIGN

Aside from the pick-up lens, view finder (references 31 and 32), and other optical requirements, the electrical and mechanical design of the studio camera must be given careful consideration. Experience has shown that the following principles should be followed in the design:

(a) In order to keep the camera as small, light, and mobile as possible, the number of tubes and components housed in the camera proper should be kept to the minimum. As much auxiliary equipment as possible should be mounted on permanent racks in or near the control room.

(b) All electrical camera controls and adjustments which may require attention during a broadcast should be located at or within easy reach of the control-room console.

(c) The mechanical layout of the camera should facilitate rapid servicing in case of failure. Such features as replaceable plug-in type video preamplifiers will simplify maintenance of the camera.

A variety of different types of camera tubes have been used for television studio pick-up. Before 1946 the Iconoscope camera found the widest acceptance among television broadcasters. Subsequently the Image Orthicon (reference 30) with its greater light sensitivity has been found to have practical advantages for studio use. Since space does not permit a detailed discussion of all camera types, the Iconscope has been chosen in the following example since it embodies many features that are common to present television cameras.

In a typical Iconoscope studio camera design, the following items are included in the camera housing proper:

(a) Video preamplifier (five tubes) including a high-peaker stage and a cathode follower to feed a 75-ohm coaxial line.

(b) Beam blanking amplifier (double triode).

(c) Horizontal and vertical deflection coils in a yoke assembly arranged to be fed from an external source via coaxial cables.

(d) Bias lighting arrangement for improving the collection of secondary electrons from the mosaic.

(e) Filament transformer (110-volt alternating current to 6.3-volt alternating current 60 cycles).

This particular design of camera used an optical view finder and was mounted on a movable pedestal equipped with a pushbutton-controlled motor-driven elevating arrangement, a suitable tilting and panning head being provided for mounting the camera.

A typical flexible cable for connecting the camera to the rack equipment has conductors and insulation as shown in Table 1. As indicated in the table, this cable has been designed as an all-purpose cable which can be used with a variety of camera and pick-up tube types. Conductors 26 and 27 are coaxial cables, conductor group 28, 29 is a shielded balanced twisted pair video cable. Conductors 3 and 4 are used to feed high voltage to the pick-up tube. Video cable shields should be run separately from the common power supply circuit ground to avoid possible interference pick-up due to common ground return impedance.

For camera cable lengths of 100 to 150 ft maximum, it is practical to locate the deflection wave generators in the control-room racks, provided sufficiently low-impedance deflection coils are used in the camera deflection yoke. For longer cable lengths, it is advisable to locate the deflection amplifiers in the camera proper or in the camera pedestal, feeding the horizontal and vertical driving pulses or sawtooth waves out to the camera from the racks via terminated 75-ohm coaxial cables or terminated balanced pair twinax cables.

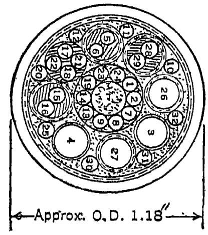
Some types of camera tubes are easily damaged if the deflection circuits fail during operation. This is true of Iconoscopes. Orthicons may also be damaged if the mosaic is illuminated when the deflection failure occurs. It is, therefore, necessary to provide protective devices to bias off the scanning beam automatically when deflection fails. Relays actuated by rectified deflection currents have proved satisfactory for this purpose.

The video output of the camera may be fed to the control room from the camera either by a coaxial line or by a balanced twisted pair transmission line with an external grounded shield. The latter is preferred to give better discrimination against hum pick-up in long cables.

The camera preamplifier design usually provides for a peak-to-peak video output from the camera of between 0.1 and 1.0 volt. The higher levels are preferred to provide discrimination against possible interference pick-up in the cable.

VIDEO PREAMPLIFIER DESIGN. The scene brightness, the speed of the lens, the sensitivity of the pick-up tube, and the value of pick-up tube signal load resistor being known, the low-frequency gain required of the preamplifier to bring the signal up to the desired level to feed the 75-ohm cable can readily be calculated. (By low-frequency gain is meant the gain for video frequencies below 50 or 100 kc.) The high-frequency gain required of the amplifier, assuming that the general practice of feeding a "flat" signal to the line is followed, is dependent on the frequency and phase distortion introduced at the input of the preamplifier by the network used to couple the output of the pick-up tube to the grid of the first stage of the preamplifier. Use of a high value of load resistor (30,000 to 100,000 ohms) compared to the value which would be chosen for flat response (1000 to

Table 1. Flexible Cable Used between Television Cameras and Equipment Racks



Conductor No.	Conductor Color	Cable Mfrs. Maximum E (AC)	Cable Mfrs. Maximum E (DC)	Cable Mfrs. Maximum I (Amps)	A.W.G. No.	Function: Orthicon	Function: Iconoscope	Function: Image Orthicon
1	White	250	450	2.4	20	Mult. focus +150 volts to +250 volts	Ike focus −350 volts to −600 volts	Mult. focus +150 volts to +250 volts
2	Black					Ring −50 volts to +150 volts		Decelerator −50 volts to +150 volts
3	Yellow	1000	1400				Ike grid −950 volts to −1150 volts	Image focus 200 volts to −1000 volts
4	Red	2000	2800			Multiplier +1000 volts to +2000 volts	Ike cathode −1000 volts to −1200 volts	Multiplier +1000 volts to +2000 volts
5	Black			4.0	18	110 volts a.c.	110 volts a.c.	110 volts a.c.
6	White							
7	Red	250	450	2.4	20	Bias −105 volts	Bias −105 volts	Bias −105 volts
8	Brown					Video +B +280 volts	Video +B +280 volts	Video +B +280 volts
9	Orange					Deflection + B		Deflection + B
10	White	50	70			Horizontal center	Edge light	Horiz. center
11	Black						Back light	
12	Red					Target −15 volts to +15 volts		Target −15 volts to +15 volts
13	Green	250	450			Beam 0 to −45 volts		Beam 0 to −45 volts
14	Black					Signal light	Signal light	Signal light
15						Vertical deflection	Vertical deflection	Vertical deflection
16	Red							
17	White					Program phone	Program phone	Program phone
18	Red							
19	Blue					Wall focus +100 volts to +250 volts		Wall focus +100 volts to +250 volts
20	Green	50	70			Alignment field		Alignment field
21	Black	250	450			Phone	Phone	Phone
22	Green							
23	White					Front focus field		Front focus field
24	Yellow							
25	Blue	50	70					
26	Brown	600	850			Horizontal S.T.	Horiz. deflect.	Horiz. driving
27	Orange					Blanking	Blanking	Blanking
28	Yellow	300	500			Video	Video	Video
29	Blue							
30	Yellow	50	70			Rear focus field		Rear focus field
31	Brown							
32	Orange					Alignment field		Alignment field

3000 ohms) offers advantages in discrimination against low-frequency noise and microphonic disturbances in the first preamplifier stage, because it results in a relatively greater level of low-frequency signal being applied to the first grid than if a flat system is used. On the other hand, tubes having a self-contained electron signal multiplier such as the Image Dissector or the Image Orthicon can provide a relatively large signal output current so that a relatively low value of signal load resistor can deliver sufficient signal voltage, in many cases, to override noise and microphonic disturbances generated in the first thermionic amplifier stage.

When a high value of signal load resistor is used, it is the practice to compensate for the frequency and phase distortion introduced by incorporating a "high-peaker" stage in the video preamplifier. Figure 2 is a block diagram illustrating one form of high-peaking.

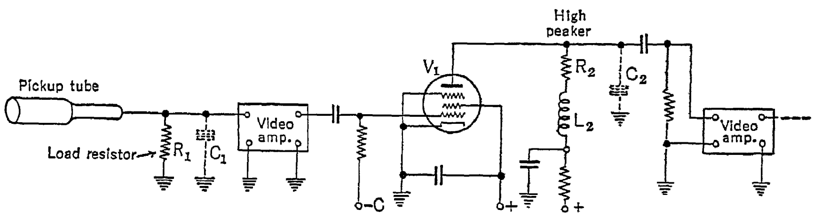


FIG. 2. High Peaker Stage Employing Resistance and Inductance in the Plate Circuit

The load impedance Z_1 for the pick-up tube consists of R_1 and C_1 in parallel, C_1 being the combined output capacitance of the pick-up tube, stray and wiring capacitances, and the input capacitance of the first video amplifier stage. The video amplifier stages following the input circuit are designed for uniform response over the required band width. These feed a high-peaker stage V_1 having an effective plate load impedance Z_2 at high frequencies consisting of R_2 and L_2 in series (neglecting C_2). Since the internal impedance of both the pick-up tube and V_1 will be relatively high (they may be considered as constant-current generators) and since the amplifier between the pick-up tube and V_1 is designed for uniform response, the overall response of the system is proportional to the product Z_1Z_2. If the values of Z_1 and Z_2 are chosen so that $Z_1Z_2 = A^2 =$ constant (inverse networks), an

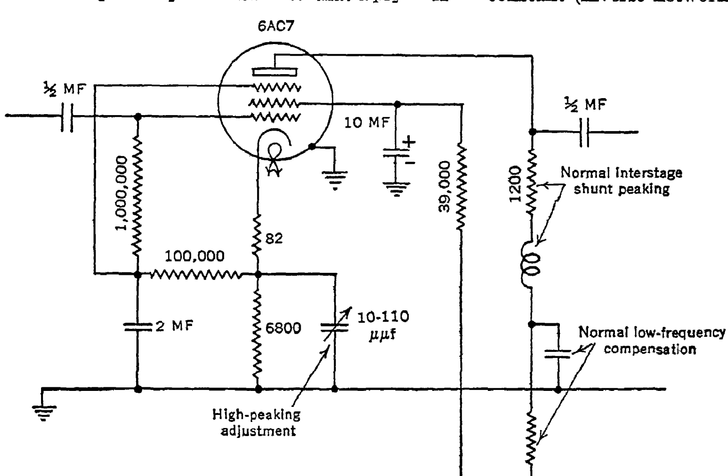


FIG. 3. High Peaker Stage Utilizing Variable Cathode By-pass

overall flat response will be obtained for frequencies below the point where the resonance of C_2 with L_2 begins to have important effect. This condition is fulfilled when $R_1C_1 = L_2/R_2$. It has been found that, if L_2 is chosen such that the resonance of C_2 with L_2 occurs at a frequency about 1.5 or more times higher than the top video frequency for which the balance of the amplifier is designed, the effect of C_2 can be neglected.

Obviously, by adding a series coil to the input circuit, a pair of inverse networks could be provided which would exactly compensate for the distortion including the effect of C_2.

From a practical standpoint, however, the simpler arrangement of Fig. 2 has been found to give satisfactory results. Two other high-peaker circuits which are simple and yet can be designed to give satisfactory compensation over a reasonably wide band are illustrated in Figs. 3 and 4. A variety of other circuits and compensation methods can also be used.

The low-frequency compensation of the preamplifier generally follows standard practice, the amplifier being designed to pass without appreciable distortion a 60-cycle square wave applied to the input terminals. In some designs of Iconoscope preamplifiers, an additional network consisting of 150,000 to 300,000 ohms shunted by 0.001 to 0.005 μf is placed in series with the low-potential end of the signal load resistor to obtain improved operation

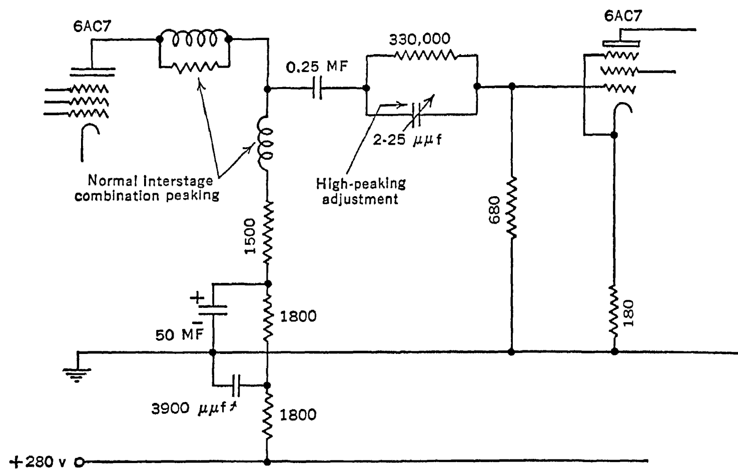


FIG. 4. High Peaker with Variable Series Capacitor

at low and medium frequencies. When this is done, an appropriate network must be inserted in one of the later stages of the amplifying system to compensate for this low-frequency pre-emphasis.

13. STUDIO EQUIPMENT

CONTROL ROOM. In a typical studio control room for Iconoscope cameras, the rack equipment for each camera chain consists of the following:

(*a*) Horizontal deflection sawtooth wave generator, with the output modulated by sawtooth waves of vertical frequency for keystoning. The amplitude of the sawtooth wave, as well as the amount of keystoning, is adjustable. An output transformer feeds the low-impedance deflection coils in the camera via a 75-ohm coaxial cable. An impedance step-up transformer located in the camera is sometimes used to feed the deflection coils.

(*b*) Vertical deflection sawtooth wave generator with an output transformer feeding a 75-ohm coaxial line which connects to the camera vertical deflection coils. An impedance step-up transformer located in the camera is sometimes used to feed the deflection coils.

(*c*) Regulated power supplies for deflection amplifiers, video amplifiers, and the Iconoscope.

(*d*) Studio amplifier. This amplifier provides video outputs for feeding monitoring circuits and the master control room. The output circuits feed 75-ohm coaxial lines at a signal voltage level between 0.5 volt and 1.0 volt peak to peak, depending on the standard level adopted for a given plant. The output of the studio amplifier contains picture blanking signals but not synchronizing signals. An excess of picture blanking signals is introduced in one of the later stages of the studio amplifier followed by an adjustable clipper or pedestal (brightness) control. The amplifier is provided with a video gain control.

(*e*) Shading amplifier. Certain types of camera tubes, including the Iconoscope, generate a spurious signal output in addition to the signals representative of picture information. The effect on the television picture of uncompensated spurious signals is to produce shaded areas not present in the original scene.

It has been found in practice that the effect of spurious signals can readily be compensated by the addition of a few simple wave forms of appropriate amplitude and polarity to the camera video output. Since the shading wave forms and amplitudes required vary with scene illumination, it is the practice to provide individual manual adjustment of shading waveforms for each Iconoscope camera. Pick-up tubes which do not always require shading signals are the Image Dissector, the RCA type 1840 Orthicon, and the Image Orthicon.

The following shading wave forms, adjustable in amplitude and polarity, are provided for each camera chain: (1) horizontal sawtooth; (2) vertical sawtooth; (3) horizontal parabola (adjustable clipping); (4) vertical parabola (adjustable clipping); (5) vertical sine wave with adjustable phase (the inclusion of vertical sine wave is optional).

When a studio is equipped with more than one Iconoscope camera, it is usually economical to provide a shading generator having a low-impedance pushpull output bus for each shading wave form. Relatively high-impedance center-tapped potentiometers can then be connected in parallel with these buses to feed a shading isolation and mixing amplifier, associated with each camera chain.

Shading signals should be added to the video signal in a low-level stage of the video system prior to the introduction of picture blanking. Introduction of shading into a low-level stage (it is sometimes fed into an early stage of the preamplifier in the camera) tends to prevent overload of the higher-level stages due to uncompensated tilt in the video signal wave form.

The following operating controls should be located on the studio console for each camera chain: (1) Iconoscope beam current; (2) Iconoscope focus; (3) video gain (contrast); (4) pedestal (brightness); (5) all shading controls.

The following controls which generally require attention only during initial warm-up of the equipment need not be mounted on the console but should be within easy reach of the video operator: (1) vertical deflection amplitude; (2) vertical deflection centering; (3) keystone adjustment; (4) horizontal deflection amplitude; (5) horizontal deflection centering; (6) bias light adjustment.

Though the example given above deals with Iconoscope cameras, the same general principles and layout apply for other types of pick-up tubes, except, of course, that different arrangements of operating controls and adjustments are necessary. The need for shading signals is generally associated with Iconoscopes alone, but availability of at least sawtooth and parabola wave forms is often advantageous when using other types of camera tubes which nominally do not require it. Thus a tube with somewhat non-uniform sensitivity over different parts of its photoelement may produce a satisfactory picture if shading signals are available.

AMPLIFIER DESIGN. In the usual equipment layout, the studio amplifier performs the following functions:

1. Provides for the insertion of shading wave forms if they have not previously been added in the preamplifier.
2. Provides for the addition of picture blanking signals of adjustable amplitude to the video signal.
3. Provides for variable gain control (automatic gain control is sometimes used) of the video signal.
4. May include one or more of the following corrective features: (*a*) gamma correction; (*b*) aperture distortion correction; (*c*) correction for phase distortion.
5. The studio amplifier usually provides at least two independent outputs of combined video and blanking signals at levels of 0.5 volt to 1.0 volt peak to peak at 75-ohm impedance. One of these outputs normally feeds the master control and switching point where the studio output may be switched to the transmitter. The other output feeds the local studio picture monitoring circuits.

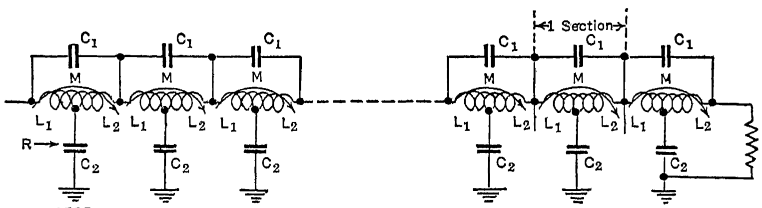
The video coupling networks used in the studio amplifier follow the conventional designs. Shading signals are generally added via a tube whose plate or cathode feeds an impedance common to a plate or cathode of one of the video stages. Blanking signals are generally added through a tube whose plate feeds an impedance common to the plate of one of the later video amplifying stages. An excess of picture blanking signal is added, being clipped off by a variable clipping arrangement in the following video stage. Care should be exercised in the design to provide an exceptionally linear clipping arrangement to avoid crowding of signal voltages corresponding to the darker portions of the picture.

MONITORING AND SWITCHING FACILITIES. Studio monitoring facilities generally consist of a picture monitor showing the outgoing picture plus one picture preview monitor for each camera in operation in the studio. The outgoing picture or "on the air" monitor should preferably be fed through a return feed from the master control point and should be synchronized from the synchronizing signals which have been added to the video and blanking components at the master control point. The preview monitors are generally of the "driven type," i.e., locked-in by the horizontal and vertical driving pulses utilized to operate the camera deflection circuits. When this is done, the preview monitors may be switched from one camera output to another without loss of synchronization. Also, since the camera outputs normally do not provide synchronizing signals, the use of driving impulses gives more stable results than blanking signals for preview monitor synchronization. The on-the-air monitor should be synchronized from pulses contained in the composite input signal, however, because it is often necessary to connect it to view pictures supplied from other sources in the normal course of operations.

One satisfactory method of switching the video monitors is to provide them with high-impedance inputs (via cathode followers in some designs) so that one or more can be switched to the 75-ohm monitoring output bus of any one of the studio amplifiers. Monitor switching may be accomplished by mechanical switches, by switching relays, or by electronic means. A relay switching system controlled by a set of pushbuttons and indicator lamps mounted on the control console is sometimes used.

In addition to picture monitoring facilities, wave-form monitors have to be provided as a continuous check on the shading and voltage level of the signal generated by each camera chain as well as that being fed out of the studio. A pushbutton-controlled switching system for the wave-form monitors is desirable.

TIME DELAY NETWORKS. In the usual television studio plant having more than one studio or signal source, it is the practice to add the synchronizing signals to the video



$L_1 = L_2 = \frac{0.158R}{f}$ = self-inductance of each coil.

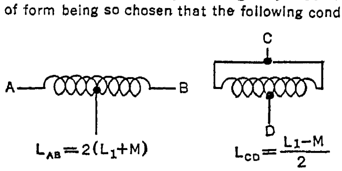
$M = \frac{0.050R}{f}$ = mutual inductance of L_1 and L_2. (Coils are wound so that the fields are series aiding, thus giving an effective negative mutual inductance).

$C_1 = \frac{0.029}{Rf}$

$C_2 = \frac{0.315}{Rf}$

f = top frequency required to reproduce original pulse.
R = design impedance.
Approximate delay per section $= \frac{0.316}{f}$ seconds.

The L_1, L_2, M assembly is a single-layer coil with a center tap, the wire size spacing between turns and diameter of form being so chosen that the following conditions are satisfied.



A center tapped coil satisfying these conditions is the equivalent of two coils each of self-inductance L_1 and of negative mutual inductance M

FIG. 5. Delay Network Employing Negative Mutual Inductance

and blanking signals at one point in the system, usually the master switching or control point. This is done for the following reasons: (1) failure of the studio output does not disturb the transmitter or throw receivers out of synchronization; (2) synchronizing signal level fed to the transmitter can be made independent of variations in studio output level; (3) insertion of synchronizing signals in the system at a single point just prior to the studio-transmitter link is the most practical way to maintain synchronizing signals at the exact level required by the transmitter.

Since the cables connecting the various studios with the master control position are usually of different lengths, electrical networks are employed to delay the signals arriving via shorter cables so that the proper "front porch" margin between the starting time of the horizontal picture blanking pulses and the starting time of the horizontal synchronizing pulses will be maintained when synchronizing signals are added. The delay networks usually consist of a number of sections of artificial line utilizing lumped constants. Though it would be possible to insert the delay networks in the studio video outputs, it is generally preferable to insert them in the coaxial lines feeding the studio with the camera driving, camera blanking, and picture blanking signals. This is more economical as less perfect delay networks can be used and the effects of any slight transmission line reflections due to incorrect termination can readily be removed by clipper stages in the studio deflection and blanking amplifiers.

DESIGN OF DELAY NETWORKS. Studio delay networks may be designed in a number of forms. A common type is an artificial line with $Z_0 = 75$ ohms in the form of a ladder network with a number of series coils and shunt capacitors. This type of artificial line may be treated as a low-pass filter with the cut-off chosen somewhat higher than the top operating frequency. As in usual filter practice, this type of network is best terminated by a double M-derived network at its output. A design using negative mutual inductance which gives excellent performance is shown in Fig. 5. The top operating frequency may be chosen on the basis of the number of harmonics of line-scanning frequency required to reproduce a pulse of a given steepness. The order n of the highest harmonic required is $n = 100/p$, where p is the time of rise of the wave expressed in percentage of the fundamental period (reference 35).

Continuous lengths of transmission line can be used in lieu of lumped delay networks. If a continuous length of line is used for time delay, it should have reasonably low attenuation and uniform time delay over the required frequency band.

14. GAMMA (TRANSFER CHARACTERISTIC)

The term gamma (γ), has been used to define the slope of the curve of the logarithm of image brightness vs. the logarithm of object brightness (reference 36). Unless otherwise specified, the slope γ refers to the central linear portion of the curve, between the extremities of the brightness values considered.

Though the term gamma is a useful concept when making comparisons with photography, the trend is away from it in connection with television systems. As a matter of fact, the term must be applied with caution since it often leads to erroneous conclusions. The reason for this is twofold. First, gamma, being a numeric and referring only to the central portion of the characteristic, tells nothing about the effect of the shape of the toe and knee of the curve; second, the concept is difficult to apply in many portions of the television system where a reference to picture black is not directly available. Therefore, the term gamma, if used at all, should be restricted to comparisons of original scene brightnesses and final image brightnesses in an overall sense.

"Transfer characteristic" (reference 28) is the name that has been given to the logarithmic plot of light input vs. signal voltage output of a television transducer. Thus, the Iconoscope performance curve of Fig. 1, plotted to logarithmic coordinates, could be called the "signal current vs. illumination transfer characteristic" of an Iconoscope.

GAMMA CORRECTION. It can be shown that the overall signal-to-noise ratio of a television pick-up, transmitting, and receiving system is improved by transmitting a logarithmic light vs. voltage characteristic. Though the exact shape of the curve has not been standardized (1949), it appears to be accepted that the light vs voltage characteristic of the Iconoscope (see Fig. 1) is close to the ideal shape. The reproducer normally has the opposite curvature to Fig. 1. This is conveniently obtained by providing a picture tube with the required light output vs. grid voltage transfer characteristic.

When a pick-up tube such as an Image Dissector or an RCA type 1840 Orthicon, having linear voltage output vs. illumination transfer characteristic, is used, it is necessary to incorporate a non-linear element in the amplifying system to achieve the desired logarithmic curvature. An amplifier designed for this purpose is called a gamma-correction amplifier.

A wide variety of circuits may be used for gamma correction. The following design principles apply: (*a*) Black level must be established at the input to correction stage so that the correction can be applied independent of the excursions of the *AC* axis of the video signal with respect to black level. (*b*) The correction should be applied prior to the video stage in which picture blanking pedestals are inserted. (*c*) The correction amplifier should be designed so that the amount of correction applied is variable between no correction at all and the maximum value. The maximum value is limited by the minimum signal-to-noise ratio that can be tolerated. A difference in video gain between the signals corresponding to the dark portions and the light portions of the picture of 10 : 1 or even 100 : 1 is often desirable. A gain difference of 10 times is the most that can be used in many cases, however, owing to signal-to-noise ratio limitations. The amount of correction to be applied for a given scene is subject to artistic as well as technical requirements. (*d*) For a television system capable of reproducing only a relatively narrow contrast range, the exact shape of the gamma-correction curve is relatively unimportant as far as the eye is concerned when viewing the overall result.

A common type of gamma-correction amplifier utilizing tubes of different characteristics in parallel is shown in Fig. 6. Another system having a diode connected across the cathode resistor of a video stage is shown in Fig. 7.

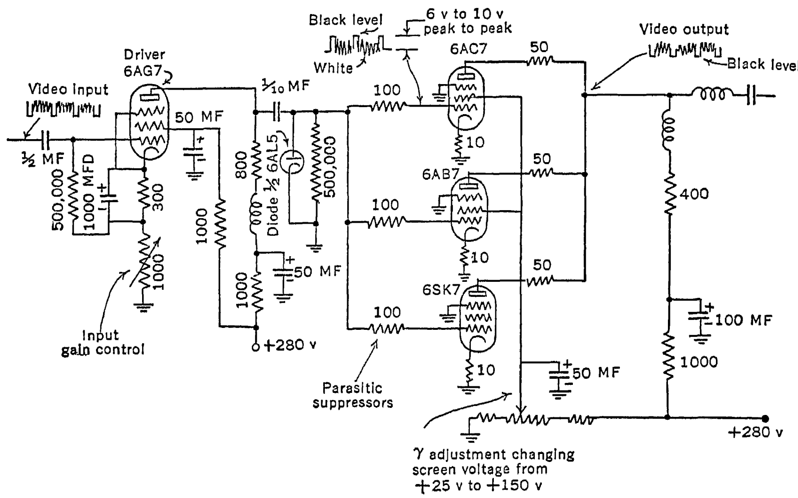


FIG. 6. Gamma Correction Stage Using Tubes in Parallel

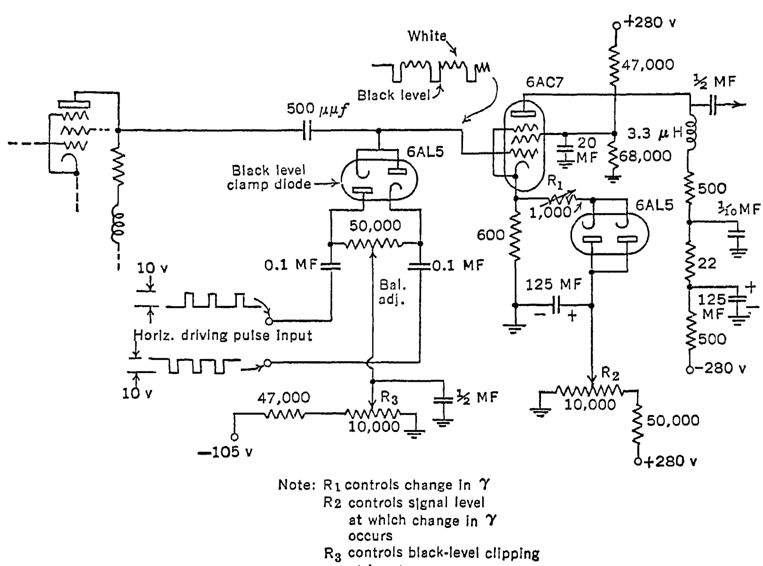


FIG. 7. Gamma Correction Stage with Variable Cathode Degeneration

15. APERTURE CORRECTION

When the size of the pick-up or reproducer scanning spot is appreciable compared to a picture element, aperture distortion takes place. For a mechanical scanning system or an equivalent system in which the dimensions of the scanning spot are accurately known, the amplitude vs. frequency distortion caused by the finite size of the aperture may be accurately calculated (reference 6). For cathode-ray systems the effects of aperture distortion are best determined by test. In a pick-up tube using cathode-ray beam scanning, for example, the effects of aperture distortion may be measured by determining the signal output obtained by scanning different patterns of uniformly illuminated alternate black and white lines of various pitches. A measured signal output vs. picture element size curve having been obtained for a given type pick-up tube, an amplitude correction equalizing network may be designed to compensate for the aperture distortion (reference 2). The complete aperture correction equalizer should include a phase equalizer to compensate for any phase distortion introduced by the amplitude correction network. Apertures of irregular shape may also introduce phase distortion. The aperture correction network, if designed for a characteristic impedance of 75 ohms, may be connected between the studio amplifier output and the coaxial line feeding the master control position.

The amount of aperture correction that can be applied in a given case is limited by the extent to which a corresponding reduction in the signal-to-noise ratio of the studio output can be tolerated. As a practical matter, it is found that the inherent resolution capabilities of most types of television pick-up tubes is great enough so that a special network for aperture correction is unnecessary. Adjustment of the high-peaker stage (see article 12) for the sharpest picture may often include some inadvertent correction for aperture-distortion effects.

16. FILM PICK-UP

To assure program continuity and flexibility the following may be considered the minimum equipment requirements for the television film pick-up studio: (1) two television film pick-up cameras with associated control, monitoring, and amplifying equipment; (2) two 35-mm sound-motion-picture projectors; (3) two 16-mm sound-motion-picture projectors; (4) two 35-mm slide projectors; (5) sound control and monitoring facilities.

The amplifying, control, and monitoring equipment for the film pick-up studio is essentially the same as that for the direct pick-up studio. Additional console controls required are remote projector starting and phasing. Since the synchronizing generator is generally locked-in with the local power mains, synchronous motors are used to operate the film projectors. A convenient method of phasing the projectors is to shift the phase of the synchronizing-generator 60-cycle reference voltage, thereby shifting the phase of camera scanning with respect to the projector. Mechanical phasing methods are also satisfactory.

Both storage and non-storage types of pick-up tubes are suitable for television film pick-up. Since the standard sound film frame rate is 24 per second and the standard television frame rate is 30 per second, the difference is made up by scanning consecutive frames 2 and 3 times respectively at the 60-cycle interlaced field deflection rate.

In the case of non-storage-type pick-up tubes, such as the Image Dissector, the optical image of the film or an appropriate section of it must be projected optically on the photocathode at the time scanning takes place. Intermittent film projection is not considered practical for non-storage tubes since the film pull-down time is limited to approximately 0.00117 second according to present standards. Special projection methods are therefore required. Some of these employ uniform motion of the film in conjunction with optical and electronic means to provide scanning of alternate film frames in the required 2,3 sequence (reference 37).

For storage-type pick-up tubes, such as the Iconoscope, the optical image is flashed on the mosaic during vertical blanking time and the stored "charge-image" scanned off during the normal vertical scanning time (reference 38). The film is moved during the vertical scanning time when the mosaic is dark. For 35-mm film a special intermittent motion designed so that the interval between pull-downs alternates between $^1/_{20}$ and $^1/_{30}$ second may be used to attain the required projection sequence, the maximum film pull-down time available being approximately 0.015 second under 1946 standards. Projectors for 16-mm film can be designed for a pull-down time of 0.007 second or less, in which case a normal equally spaced intermittent may be used. A synchronously driven rotating shutter placed between the projector lens and the pick-up tube is used to flash the picture on the mosaic. Proper phasing between the light flashes and the scanning may be obtained by mechanical or electrical means.

Automatic brightness control (automatic adjustment of the picture blanking clipper bias in the studio amplifier) may be provided for film pick-up by means of an auxiliary photocell to pick up the light from the film a few frames ahead of projection. This photocell output is integrated and applied to the blanking clipper bias circuit.

17. MASTER CONTROL POSITION

In the usual television plant, where there are several signal sources, including remote pick-ups, all switching and the final monitoring of the signal before it is sent to the transmitter are done at a central point called the master control position. The synchronizing generator, the synchronizing signal distribution amplifiers, the video line amplifiers, and the various items of test equipment used in connection with operations are usually located in a main equipment room adjacent to the master control position.

A typical studio master control pulse and video system layout is shown in Fig. 8. The synchronizing generator outputs are fed to the inputs of a number of pulse distribution

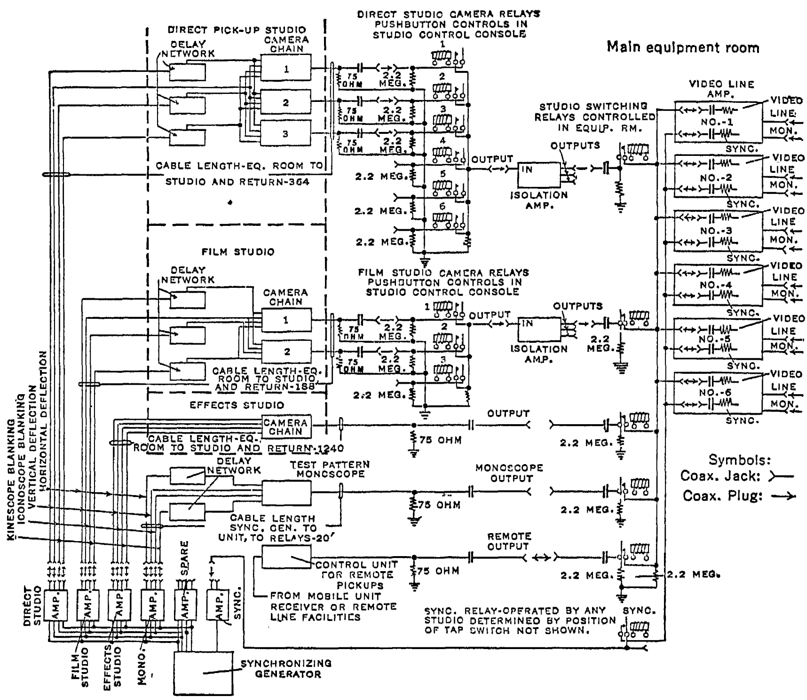


FIG. 8. Block Diagram Showing Pulse Distribution to Studios and Video Switching Circuits at the Master Control Position

amplifiers as shown. A separate pulse distribution amplifier feeds each studio and the monoscope. A separate distribution amplifier is also reserved for synchronizing signals. This system of pulse distribution isolates one studio from another so that an accidental short circuit or failure in a pulse line to one studio cannot affect the others. A spare pulse distribution amplifier is held in readiness in case of failure of one of the amplifiers. It is also good practice to have a spare synchronizing generator in operation at all times and to provide means for quickly switching it into service if the regular generator fails.

As shown in the diagram, delay networks are inserted in the pulse lines to all studios except the one having the longest cable run. The delay networks are adjusted (by adding or removing sections) so that the time of arrival of the blanking signals from each studio at the master control position switching relays will be the same and will be such that the proper front porch delay of synchronizing signals will be maintained when the signals are added. There is generally enough coincidental time delay in the camera vertical deflection system so that no delay networks are required to time the vertical pulses properly. It will be noted that the output of each studio camera is brought to a bank of switching relays associated with that studio and located at the master control position. These relay switching banks in turn feed isolation amplifiers which are arranged to feed banks of relays connecting the studio outputs to the line amplifier inputs. Each line amplifier provides two outputs, one of which can be patched to feed the transmitter and the other can provide a monitoring return-feed to one of the studios.

Normally, only one or two of the amplifiers (regular and spare) will be used for feeding the picture transmitter. The other line amplifiers are patched to provide monitoring feeds for the studio control rooms, announcers' booths, clients' booths, or other points in the television plant where line-fed monitors are used. As indicated, coaxial patch cords in connection with coaxial jacks are provided so that changes in system interconnections can be rapidly effected as dictated by operating requirements. In addition, a number of spare coaxial cables connecting to each studio terminate in a jack-field adjacent to the master control console. This makes it possible to patch up special circuit arrangements as required. Means are also provided for patching up special control circuits. For example, the effects studio output, which may be a picture of a title slide, can be patched into one of the spare camera relays in the relay bank associated with the direct pick-up studio. Thus the video operator in the direct pick-up studio merely presses a camera switching button when he wishes to insert the slide in the outgoing program.

The master control position is equipped with picture and wave-form monitoring facilities, a television receiver used as an off-the-air monitor, and audio monitoring equipment. The master control desk is normally equipped with a pushbutton-controlled monitoring system so that the incoming picture from any of the sources can be checked as required. A wave-form monitor associated with the picture monitor is used for checking signal levels, pulse wave forms, etc. All program switching between the various signal sources and the transmitter is done at the master control position.

A control unit for handling remote pick-ups consisting of a picture monitor, a wave-form monitor, and a variable gain video amplifier is usually provided. The output of this unit feeds into one of the relays in the line amplifier bank as shown. A relay is also provided to open the line amplifier synchronizing signal input when the remote relay closes, as the remote signal normally arrives complete with synchronizing signals. When the remote synchronizing generator is locked-in with the same power supply as the main generator, it is possible to adjust the phase of the two synchronizing signals so that the rapid switch from local to remote pick-up causes no appreciable disturbance of receiver scanning.

Conventional relay interlock systems are used to prevent closing of more than one relay in a particular bank at any time and to drop out one relay when the next is picked up. Magnetic delay circuits can be arranged to delay the drop-out of one relay until just before the next closes.

A wide variety of electronic fading and switching systems may be used in addition to the relay system shown. Fades, lap dissolves, superposition of one image on another, in whole or in part, may be accomplished electronically.

SYNCHRONIZING GENERATOR. Electronic synchronizing signal generators, rather than electromechanical generators, are almost universally used for providing the pulses needed to operate the television plant. The type generally used (reference 27) incorporates an AFC oscillator operating at double line frequency in connection with a frequency divider having an output at field frequency. The output of a discriminator comparing the local 60-cycle power frequency with the output of the frequency divider controls the AFC oscillator. By means of a number of multivibrators, delay networks, and pulse shaping networks, the generator provides the following outputs: (1) synchronizing signals; (2) line-frequency driving pulses; (3) field-frequency driving pulses; (4) camera blanking signals; (5) picture blanking signals. Controls are provided for regulating the duration times or "widths" of the various pulses. The relative starting times of the various pulses are usually controlled by means of an electrical delay network incorporated within the unit. The outputs of the synchronizing generator are normally available at peak-to-peak voltages of 4 to 6 volts at an impedance of 75 ohms.

The general practice has been to use the local 60-cycle power source for locking-in the synchronizing generator. However, some designs incorporate a selector switch for locking the synchronizing generator in either with the local power frequency or with a subharmonic of a crystal oscillator. Operation of the synchronizing generator from a stabilized frequency source independent of the local 60-cycle power mains has many advantages.

18. PULSE MEASUREMENTS

In order to insure that the transmitted blanking and synchronizing pulse wave forms conform to the FCC standards, the television broadcasting plant must be provided with test equipment for measurement of the relative starting times, slopes, and duration times of these pulses.

Space does not permit a discussion of the great variety of measurement techniques that can be or have been successfully used. However, the following have been found to be convenient and are sufficiently accurate for practical purposes (reference 35):

SINE-WAVE SWEEP. A method that allows rapid and accurate checks to be made of pulse slopes, widths, and delay times utilizes an oscilloscope with sine-wave sweep of

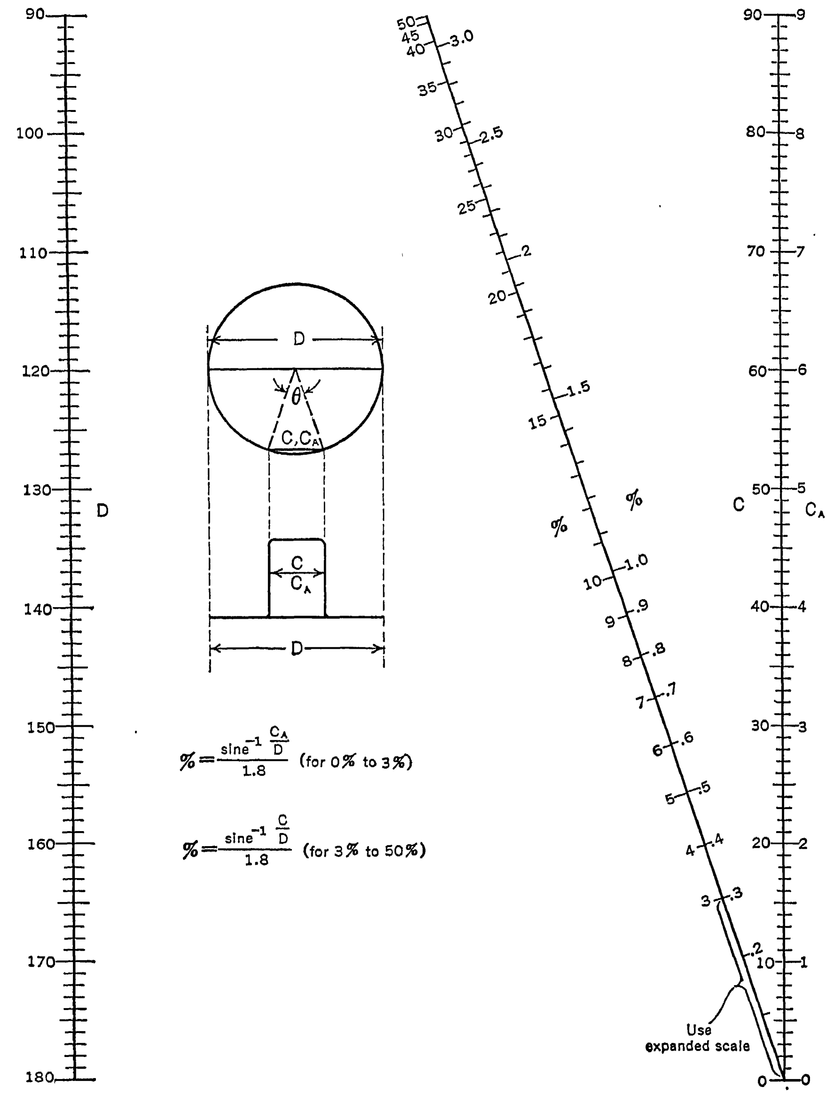


FIG. 9. Alignment Chart for Pulse Width, Scope, and Relative Time Delay Measurements Using Sine Wave Horizontal Sweep

either line or field scanning frequency depending on the character of the pulse to be measured (reference 35). The line-frequency sine wave is obtained by filtering out the fundamental of the horizontal pulses from either the synchronizing signal, the picture or camera blanking signals, or the horizontal driving pulses. For accurate results, the sine wave used for measurement purposes should be well filtered so that the arithmetic sum of all harmonics does not exceed 1 per cent. The local 60-cycle a-c mains, filtered as required, may be used as a source of field-frequency sine wave if the field frequency is locked in with the local power source. The oscilloscope used should have a linear horizontal sweep width of about 100 mm minimum. Pulse widths are measured by shifting the pulse to the center of the screen by means of a sweep phase shifter and measuring the dimensions indicated on Fig. 9 with a transparent millimeter scale at the 10 and 90 per cent amplitude points of the wave.

The width of the pulse in percentage of a sine wave period may then be obtained from the nomographic chart. The expanded scale should be used for widths of less than 3 per cent. Pulse slopes are measured by shifting the pulse edge in question to the center of the screen so that the 10 and 90 per cent amplitude points are symmetrically disposed about the center line of the sweep and scaling C or C_A and D. The C dimension appears magnified compared to linear sweep when sine wave is used, and the accuracy of measurement is thereby enhanced. Pulse delay times may be measured with respect to a specific time (such as the starting time of horizontal blanking) by shifting the phase of the sweep until the starting time of horizontal blanking and the starting time of the pulse in question appear symmetrically disposed about the center line of the horizontal sweep. C and D may then be scaled off and the relative delay in percentage of a horizontal scanning period read from the chart. The delay time of a network may be similarly measured by comparing the difference in position on the sweep of an input and output test pulse.

A variation of the sine-wave method uses an accurate phase shifter calibrated in percentage of a scanning period. This does away with chart and millimeter scales, and the linearity of horizontal sweep of the oscilloscope is no longer a factor as slopes and widths may be read off by shifting the wave known amounts with respect to a fine vertical line drawn down the center of the oscilloscope screen.

PULSE CROSS METHOD. The number of vertical synchronizing signal sections, the number of equalizing impulses before and after the vertical synchronizing impulse, and the approximate widths of the synchronizing and blanking signals may be determined by the pulse cross method (reference 35). These determinations are accomplished by locking in the picture monitor with horizontal and vertical pulses which have been delayed half a period and reversing the polarity of the composite video input signal. The pulse and blanking signals then appear as a white cross in the center of the picture tube screen. The items mentioned above may then be determined by measurement and observation of this pattern. The vertical deflection amplitude should be expanded by about 3 : 1 while making this test. A simple switching arrangement can be arranged to shift the picture monitor from normal to pulse cross operation rapidly during operation.

19. OVERALL VIDEO SYSTEM RESPONSE

In the usual television plant the picture signals pass through a relatively large number of amplifier stages in cascade in traveling from the camera to the transmitter. The transient response of the overall system must therefore be given careful consideration. A small phase or amplitude distortion in each individual stage has a cumulative effect when a large number of stages is operated in cascade (reference 39). The overall effect of such distortion when not compensated is to cause transients of an oscillatory nature to occur whenever the scanning spot encounters an abrupt change in scene brightness.

Mathematical analysis of the overall transient characteristics of a practical system is not only difficult but is complicated by the fact that all stages will not ordinarily use identical forms of high-frequency compensation or peaking. A practical engineering approach to the problem which yields a satisfactory solution is the following: (*a*) When laying out the plant, an accurate estimate can be made of the number of stages likely to be connected in cascade. This estimate can be used in conjunction with published data (reference 39) to decide on reasonable values of design parameters for high-frequency compensation of individual stages. (*b*) The design parameters for high-frequency video compensation having been chosen, the time delay distortion, due to a number of stages in cascade, can be calculated. If the difference in transmission time between medium frequencies (100 to 200 kc) and high frequencies (5 Mc) for the number of stages in cascade begins to approach an appreciable fraction of the time of one picture element, then it is advisable either to choose other design parameters giving smaller time delay distortion per stage or to employ a properly designed phase compensation network. (*c*) It is advisable to choose video amplifier design parameters such that the overall video amplitude response will be uniform within ±1 db. This can be accomplished by choosing a top video frequency (for design purposes) somewhat higher than the nominal top frequency handled by the transmitter. Thus, the studio equipment amplifiers might be designed and compensated for a top frequency of 6 Mc to assure uniform overall amplitude characteristics up to the 4.5-Mc nominal top frequency dictated by present standards. When this is done, however, care should be taken that there are no appreciable peaks in the amplitude response beyond 4.5 Mc. A gradual decay in amplitude response beyond 4.5 Mc rather than an abrupt change is desirable. (*d*) The low-frequency compensation of the video system (60 cycles to 100 kc) should follow conventional principles, considering the number of amplifier stages in cascade. (*e*) Having observed the above design precautions and provided a system having uniform amplitude response (within ± 1 db) from 60 cycles to 4.5 or more megacycles, the overall transient characteristics of the system should be investigated experimentally by square-wave techniques. The low-frequency transient response may be investigated by means of 60-cycle square waves. The high-frequency transient response may be investigated using steep-sided square waves of a fundamental frequency of 100 kc (reference 42). The 100-kc square wave should be of sufficient steepness to represent harmonics of 100 kc beyond the cut-off frequency of the system. The equivalent phase and amplitude characteristics of the system may be obtained from the square-wave response wave shape either by mathematical analysis or by means of special charts (reference 40). Appropriate phase and amplitude correction networks may be designed on the basis of square-wave test data.

Standards should be set as to the maximum allowable amount of overshoot and following transients that can be tolerated in the square-wave response characteristics of a television broadcast system. In any specific case, however, one can form an opinion as to whether objectionable transient effects exist by observing the television image, preferably an image of a resolution test pattern having a number of sharply defined boundaries of high difference in contrast. An overshoot or oscillatory transient differing by 2 per cent from the final steady-state value of a square wave will normally be noticeable in the image.

20. TELEVISION FIELD PICK-UP EQUIPMENT

From the electrical standpoint, television field pick-up equipment is quite similar to studio equipment, the major circuit functions being identical. Mechanically, however, the equipment is usually segregated into small, light-weight units of suitcase style so that it may be readily transported (reference 44).

A block diagram of a typical complement of field pick-up equipment for two cameras is given in Fig. 10.

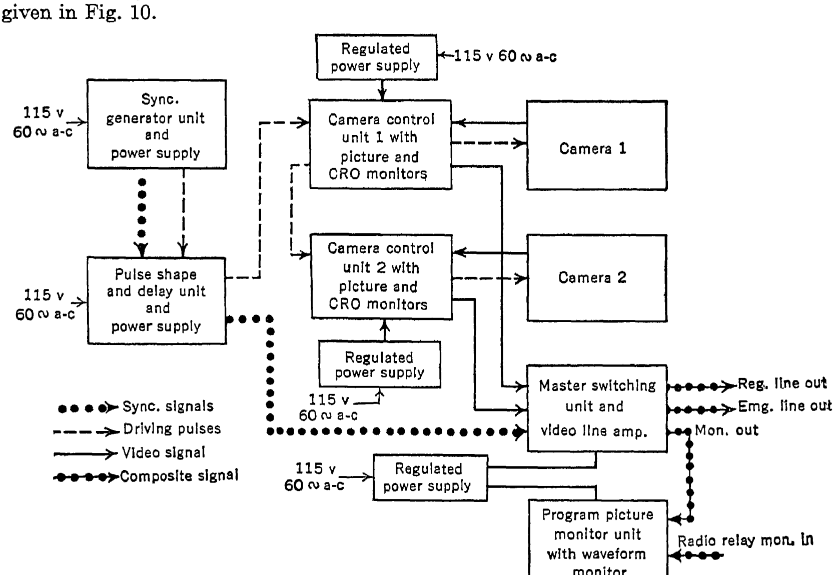


FIG. 10. Block Diagram of a Typical Complement of Portable Field Equipment for Two-camera Operation

21. RELAY OF TELEVISION SIGNAL

Television network facilities may be classified as "intercity" and "local." Under the first category are included long-distance coaxial carrier cables and radio relay multilink circuits. The latter class includes existing telephone pairs and special shielded television pairs, video coaxial cables, and short-haul microwave radio pick-up links.

INTERCITY TELEVISION FACILITIES. Coaxial cables (references 45 and 46). The long-distance coaxial cables consist of several coaxials enclosed in a lead sheath along with ordinary paper-insulated telephone pairs. A 3/8-in. longitudinal seam coaxial is common, although 0.27-in. coaxials have also been used. Figure 11 pictures a coaxial unit open at the end to show the construction. See also Section 10.

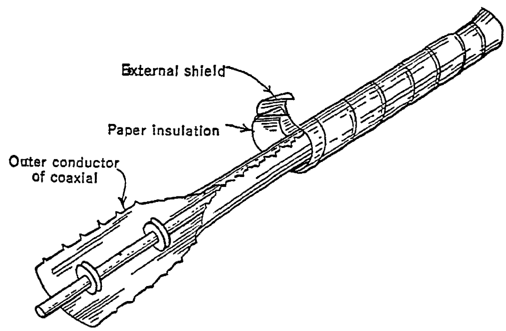


FIG. 11. Typical Construction of a Coaxial Cable Using a Minimum Amount of Solid Dielectric

Figure 12 shows the characteristic impedance, the phase delay, and the attenuation of 3/8-in. coaxials. The useful frequency range for television purposes on such a system depends upon the type and spacing of the repeaters. The present coaxial system employs repeaters at approximately 8-mile intervals and transmits television with a 311-kc carrier up to about 3.1 Mc, giving a 2.8-Mc video band. These lines must be carefully equalized for gain and delay characteristics. Figure 13 shows overall characteristics obtained (1946) on the New York–Washington coaxial system.

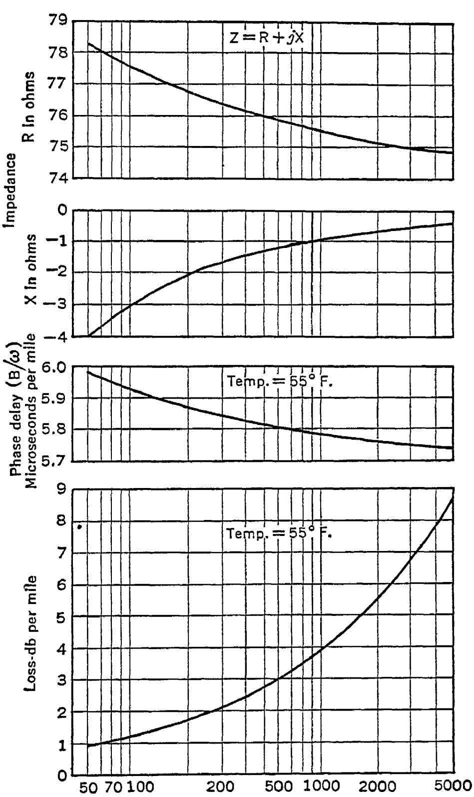


FIG. 12. Transmission Characteristics of a Coaxial Cable Constructed as in Fig. 11 (outside diameter, 3/8 inch)

Transmission of television signals over local video facilities is direct; i.e., the entire band of frequencies produced by the television camera from a few cycles per second to high frequencies is transmitted directly over the line. On long-distance coaxial facilities, however, a carrier method of transmission is used to avoid the effects of low-frequency interference. By using a relatively low carrier frequency and vestigial sideband transmission, the band-width requirements of the repeaters are not materially increased over a noncarrier system. Figure 14 illustrates a typical carrier transmission system for a 2.8-Mc pass band in which a dual modulation and demodulation process is utilized.

Only the frequencies above about 200 kc are used to transmit the television signal over the line, although the space below 200 kc may be used for the accompanying sound-program channel. The main reason for not using very low frequencies is that such coaxial systems would be noisy and difficult to equalize. In the frequency range used, coaxial systems are extremely quiet and relatively easy to equalize. For very long distances, the cost and complexity of the terminal apparatus are small compared to those of the rest of the circuit, but for short video facilities they might be objectionable.

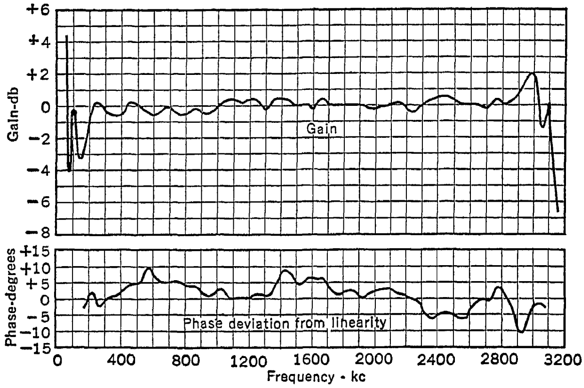


FIG. 13. Overall Characteristics of the New York–Washington Coaxial Circuit (1946)

Microwave Radio Relay (reference 47). Multilink radio relay systems also provide an important network facility for the transmission of television programs. No general rules can be given governing the choice between coaxial cable and radio relay, however, as each specific installation must be carefully studied from a number of standpoints. Among the factors governing such a choice, economic considerations will usually be paramount. Such questions as relative operating and maintenance costs and network reliability must

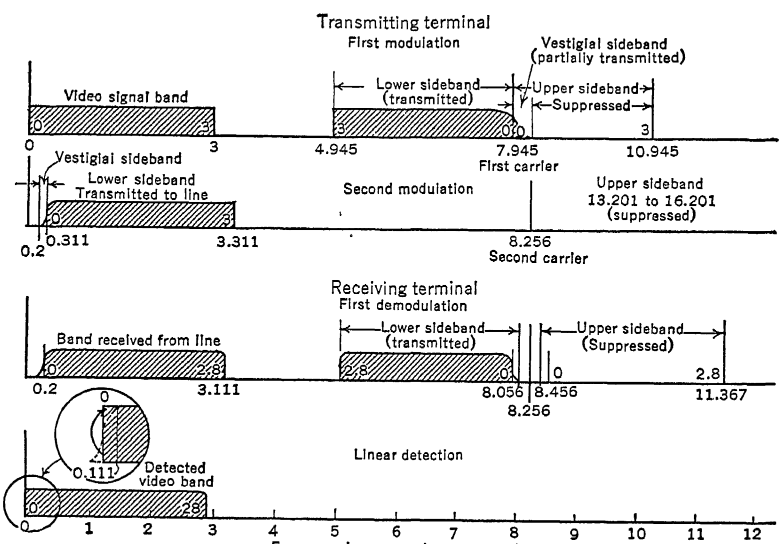


FIG. 14. Diagram of a Typical Dual Modulation and Demodulation Process Used for Transmission of a 2.8 Megacycle Video Band over a Long Coaxial Cable

also be considered for each installation. Figure 15 indicates the repeater gains required for both methods of transmission assuming a useful band width of 5 Mc. It will be noted that, except for extremely large-diameter coaxial cable, fewer repeaters are necessary for the radio relay system. This figure also illustrates the reduction in repeater gain which may be effected by employing the higher frequencies in the microwave region when a given size antenna reflector is used.

LOCAL TELEVISION CIRCUIT FACILITIES. Various local television transmission facilities may include ordinary telephone pairs, specially designed shielded pairs, coaxial cables, or microwave radio systems. Intracity wire transmission of television is accomplished at video frequencies thus avoiding the use of carrier terminals for the short distances encountered.

Telephone Pairs (references 45 and 46). Typical transmission characteristics of telephone pairs are shown in Fig. 16. Because of the large attenuation of the higher video frequencies, it is necessary to install suitable amplifiers at spacings of an average length of 1 mile, though this interval may be increased somewhat when the larger gages are employed.

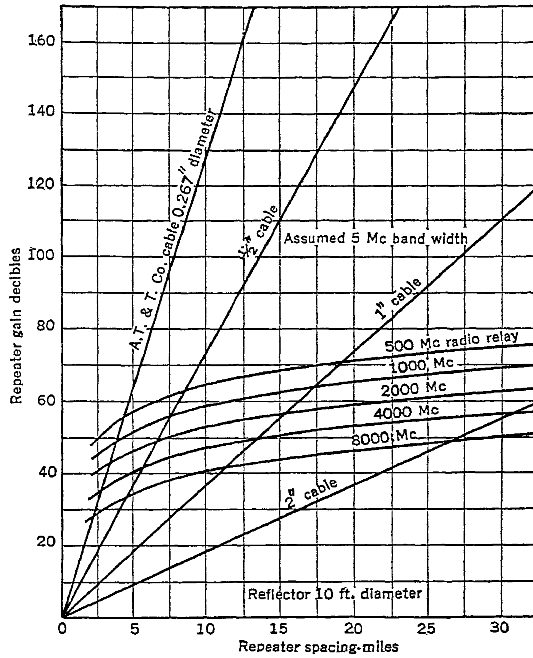


FIG. 15. Comparative Repeater Spacings and Gains for Coaxial and Microwave Circuits

It is necessary to equalize the attenuation and phase characteristic vs. frequency of the telephone pairs over the video band of frequencies. This is accomplished by means of variable equalizers associated with each video amplifier. Pre-emphasis of the level of the higher video frequencies is usually employed in these transmission circuits to obtain an improved signal-to-noise ratio.

Other factors to be considered in the use of telephone pairs for video transmission are the selection of suitable pairs within the cables and the removal of bridged taps on the pair selected. The selection of pairs within a chosen cable sheath is made with a view toward reducing interference

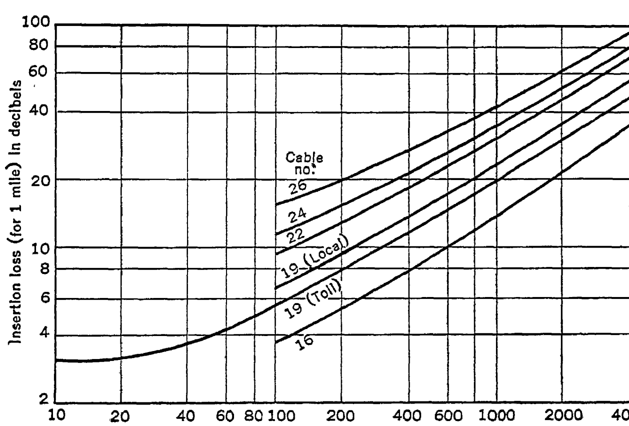


FIG. 16. Insertion Loss of Various Types of Telephone Pairs for Video Frequencies

from adjacent circuits and to avoid cross-talk coupling around the video repeater. All bridged taps are removed to assure minimum impedance irregularities due to these lumped constants along the lines.

Shielded Pairs. Where it is desirable to extend the length of the wire circuit between video repeaters, the use of special shielded pair is indicated. An opened section of such a cable is shown in Fig. 17. The transmission characteristics of such a shielded pair are

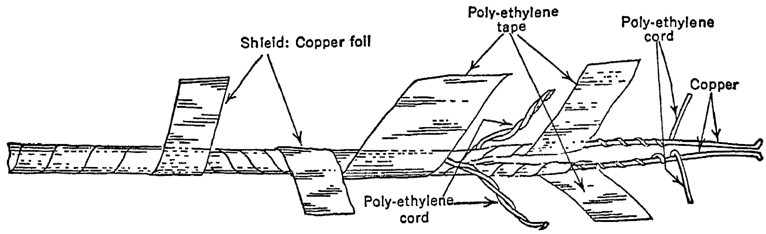


FIG. 17. Low Loss Balanced Shielded Video Transmission Line

given in Fig. 18. With this pair it is possible to extend the video repeater spacing to approximately 3.5 miles.

This type of cable consists of a pair of No. 16 gage wires insulated and spaced by means of polyethylene strings. This core is then enclosed within a cylindrical shield formed with metal tapes. A number of these units may be enclosed in a lead sheath, or the shielded pair unit may be inserted in an ordinary telephone cable replacing a certain number of the usual paper-insulated pairs.

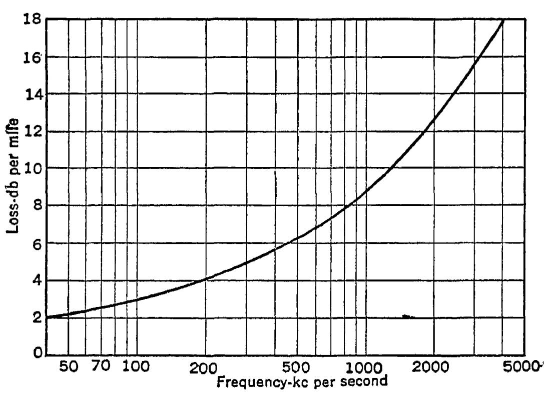


FIG. 18. Attenuation Characteristics of the Line Illustrated in Fig. 17

Coaxial Cable (references 45 and 46). Although of primary importance for network facilities, coaxial cable may be used for intracity circuits of a more permanent nature, such as studio-transmitter link service. For short distances, transmission of the video signals without utilizing a carrier system is practicable. The inherent unbalanced properties of the coaxial cable, however, usually require that special low-frequency balancing circuits be employed. A typical "hum" balancing circuit is shown in Fig. 19. Included in this figure is a simulated generator of low-frequency interference. It will be noted that noise current components flow in opposite directions through the cable terminating resistor R_1 and the hum balancing potentiometer R_2. By adjustment of the balancing potentiometer, a condition can be found such that equal and opposite noise voltages are developed between points a–b and points b–c. The noise is thereby reduced

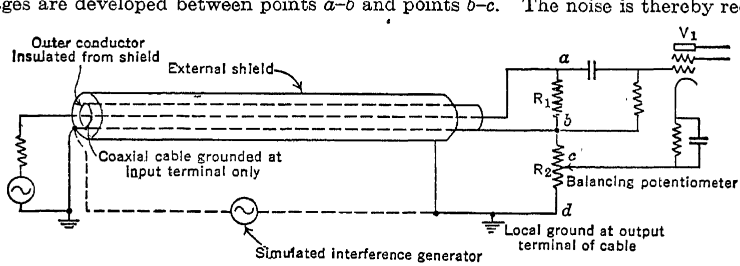


FIG. 19. Hum Balancing System for a Coaxial Cable Having an External Lead Sheath Insulated from the Copper Outer Conductor

at the input of the amplifier so that the signal voltages alone are applied between the grid and cathode of V_1.

Equalization of transmission of coaxial cables is usually accomplished by the use of an equalizing network inserted between the cable and the input of the receiving video amplifier.

In certain applications where coaxial and balanced transmission facilities are to be connected, special amplifier circuits or wide-band video repeating coils are employed to make the transition.

Microwave Radio Relay. Microwave radio relay is also an important local facility for video transmission. Many television field programs, such as parades and special events broadcasts, are not repeated at frequent enough intervals from a given location to justify the expense of wire facility installations. Furthermore, programs originating at distances greater than 20 miles from the studio plant are usually more economically handled by radio relay. Proper evaluation of these and related factors is necessary to determine the choice between radio relay or wire facilities. Radio relay may also be employed for studio-transmitter link service.

The power requirements for a microwave relay system may be approximated by the following formula. Although accurate only for the free-space propagation condition, application of this formula in practice where line of sight exists will yield results of sufficient accuracy to be useful. It should be noted that the maximum total received power due to ground reinforcement can approach, as a limit, four times the received power for free-space propagation as obtained from the formula.

The free-space transmission formula is (reference 48):

$$P_t = \frac{P_r d^2 \lambda^2}{A_r A_t} \tag{5}$$

where P_t = power fed to transmitting antenna at input terminals } Same units
P_r = power available at output terminals of receiving antenna }
A_r = effective area of receiving antenna }
A_t = effective area of transmitting antenna } Same units
d = distance between antennas }
λ = wavelength }

The power necessary at the receiving antenna output terminals depends, among other factors, upon the signal-to-noise ratio requirements of the relay system. The noise level due to thermal agitation at 20 deg cent may be computed from the following expression (reference 47):

$$P_n = (0.8 \times 10^{-20})(B) \tag{6}$$

where P_n = noise power in watts due to thermal agitation.
B = twice the highest modulation frequency in cycles per second.

In practice, the noise level due to all equipment causes will usually be between 10 and 15 db above thermal.

The effective area of an antenna is directly proportional to the power gain. The following tabulations indicate the effective areas of several typical antennas.

Antenna	Effective Area
Isotropic radiator	$\lambda^2/4\pi$
Half-wave dipole	$0.1305\lambda^2$
Parabolic reflector	Two-thirds of the projected area of the paraboloid

Several factors affect the effective area of a paraboloid, the most important being the efficiency of excitation. For example, the effective area is reduced to approximately three-eighths of the projected area when only half of the exciting antenna energy is directed toward the reflector. If transmission lines or wave guides are used in the antenna system, the attenuation due to these components should be taken into consideration when applying the free-space-transmission formula.

22. TRANSMITTER PLANT TERMINAL EQUIPMENT

Terminal equipment located at the television transmitting plant performs the function of raising the signal level delivered by the program source to that required by the transmitter and provides the necessary picture and wave-form monitoring facilities. The equipment is usually installed in a shielded room; it may consist of the following units:

A. An amplifier with means of controlling the composite signal amplitude.

B. An amplifier with independent amplitude control of the synchronizing and picture portions of the composite video signal.

C. Video switching system to select one of several sources of signal.

D. A line amplifier of sufficient output to meet the input level requirements of the video section of the transmitter.

E. Picture and wave-form monitors.

F. Monitor switching system to select circuits to be monitored.

Figure 20 shows a simplified block diagram of the visual portion of a typical television transmitter plant. Equipment required for the television sound channel follows standard frequency-modulation broadcasting practice.

VISUAL CARRIER FREQUENCY GENERATION. The r-f carrier signal for a television transmitter is generally developed by conventional methods. The primary source of radio-frequency energy is usually a highly stabilized quartz-crystal oscillator operated at a relatively low frequency. This low-frequency, low-power signal is multiplied and

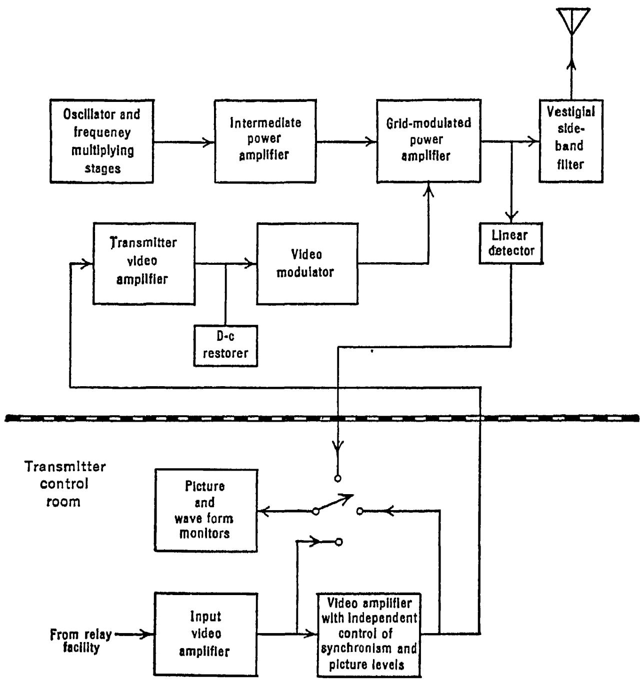


FIG. 20. Picture Transmitter Block Diagram Showing Video Input Equipment

amplified to the frequency and power level required at the modulated amplifier stage of the transmitter.

MODULATION METHOD. A few of the many possible methods of modulating the visual carrier (reference 49) are illustrated in Fig. 21. Of these, grid-bias modulation is almost universally used.

Modulation may be either at low or high r-f level. At low level the grid-bias-modulated r-f amplifier is followed by one or more class B linear r-f amplifier stages having the required band-pass characteristics.

MODULATED AMPLIFIER. The plate and grid (if grid-bias modulation is used) tank circuits of the modulated amplifier as well as all succeeding r-f stages must be capable of passing the generated sideband power without excessive amplitude or phase distortion. This requires tank circuits of relatively low impedance resulting in rather poor operating efficiencies as compared to sound transmitters.

NEUTRALIZATION. The band-pass characteristic of a television transmitter using triode tubes depends not only upon the circuit elements but also upon the effectiveness of neutralization. At low frequencies, where lead inductances may be neglected, a simple capacitance bridge adequately represents the neutralizing circuits. At higher frequencies, however, where lead inductances become appreciable, additional compensation is usually necessary, especially for wide-band operation. Figure 22 indicates the stray inductances often encountered and methods of compensation (reference 50). Stray or undesired coupling between input and output circuits not only disturbs neutralization but also affects the band-pass characteristics of the amplifier.

D-C TRANSMISSION. According to 1946 standards, the tips of the synchronizing signals correspond to maximum carrier envelope amplitude, and this is held as nearly constant as possible during a given transmission. Black level, which corresponds to the base of the synchronizing pulses, is maintained at a fixed percentage of the maximum carrier envelope amplitude within narrow tolerance independent of the values of light and shade in the picture transmitted. In order to achieve this result one or more of the various forms of "d-c restoration" circuits are used in the modulator stages of the transmitter. (See article 7.)

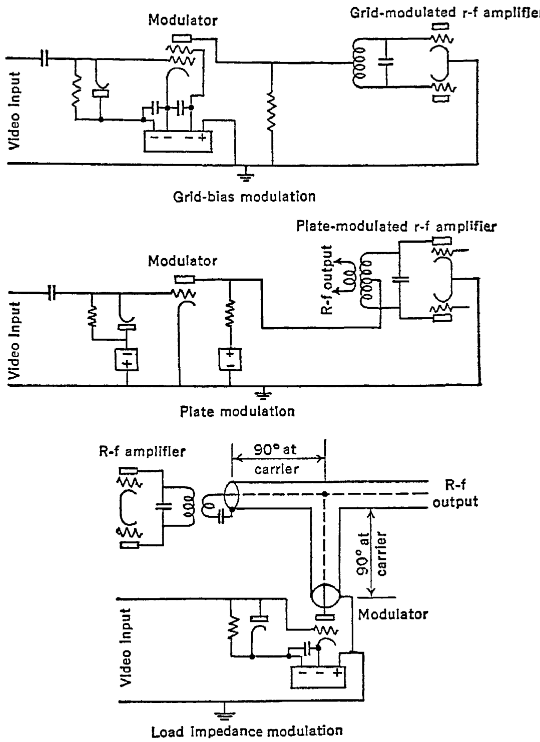


FIG. 21. Video Modulation Methods for Amplitude Modulation

VESTIGIAL SIDEBAND TRANSMISSION. A vestigial sideband system of television transmission (reference 51) is standard for commercial television broadcasting. The higher-frequency sideband components up to but not in excess of carrier plus 4.5 Mc and the lower-frequency sideband components down to carrier frequency −0.75 Mc are transmitted. The remainder of the low sideband energy must be attenuated as rapidly as possible and must reach and retain a low order of magnitude at frequencies lower than carrier frequency −1.25 Mc. Curve C of Fig. 20, article 9, illustrates the standard vestigial sideband transmission characteristic.

Transmitters which are modulated at high power level must be followed by a vestigial sideband filter which absorbs the developed but undesired sideband energy (reference 52). Vestigial sideband filters employ various configurations of elements and are usually constructed in the form of sections of concentric transmission lines of suitable lengths, diameters, and diameter ratios. A single section of one type of vestigial sideband filter is shown schematically in Fig. 23. Generally, two or more sections of this type of filter are used. In addition, "notching" filters are often required to provide additional attenuation at the low-frequency edge of the assigned channel.

Transmitters which are modulated at low power level may achieve the required transmission characteristics by proper design and adjustment of the interstage coupling networks of the following linear amplifiers. In this case, a less pretentious vestigial sideband filter may be necessary at the output of this type of transmitter in order to achieve the required characteristic.

RADIO-FREQUENCY MONITORING. Since the output of a vestigial sideband transmission system is viewed on receivers having specified band-pass characteristics, a monitor should be provided which not only conforms to the standard receiver characteristic but also yields a signal that is a true sample of the radiated energy from the transmitter.

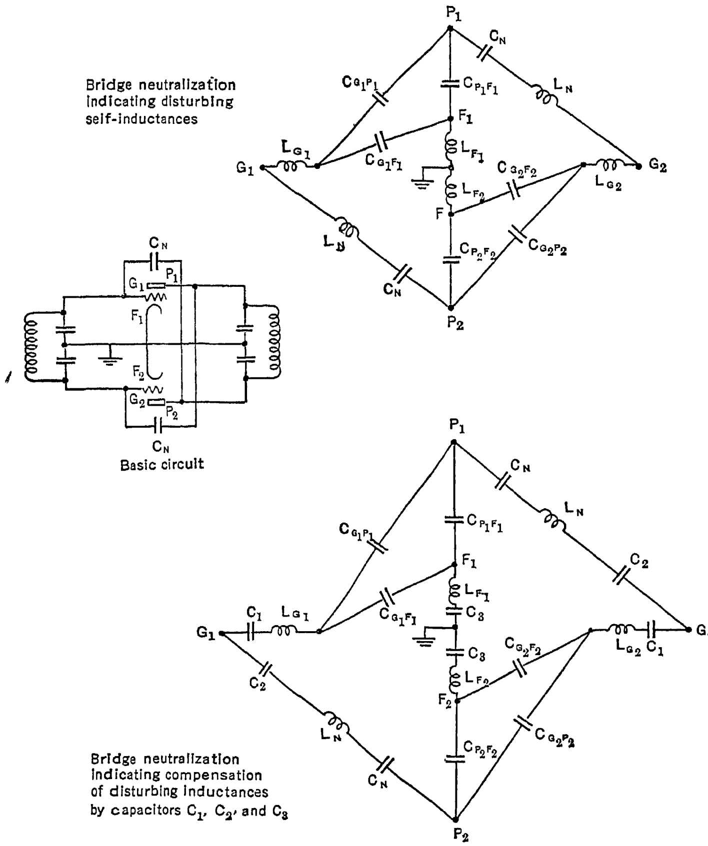


FIG. 22. Neutralization Method to Compensate for Stray Inductance

The aural transmitter may be monitored with equipment similar to that developed for the frequency-modulation broadcast service.

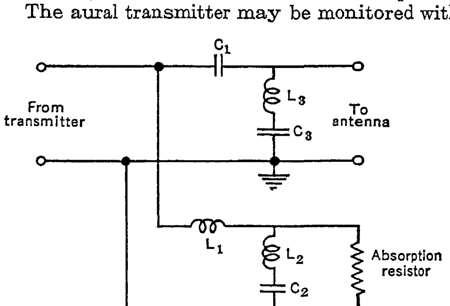


FIG. 23. Single Section Vestigial Sideband Filter

MODULATION MEASUREMENT. The modulation of the television transmitter may be measured in various ways, but the methods that take advantage of the fact that a television transmitter is operated at a constant peak carrier level have been found most satisfactory in practice.

Since the tips of the synchronizing signals represent 100 per cent modulation, one relatively simple method is to observe the carrier envelope pattern at radio frequency on an oscilloscope as shown in Fig. 24. At the high carrier frequencies involved, however, it is sometimes difficult to insure that the cathode-ray pattern is a true representation of the developed carrier envelope amplitude.

A method which avoids dealing with the radio frequency directly utilizes the output of a linear rectifier applied to the vertical plates of the oscilloscope, normal sawtooth sweep of a convenient frequency being used for horizontal deflection. A contactor is provided to short-circuit the output of the diode periodically. This provides a reference level corresponding to complete modulation in the white direction, or zero carrier envelope amplitude. The modulation percentage may be scaled off the cathode-ray screen as indicated in Fig. 24 which shows the appearance of the pattern using high frequency sweep. This simple method is quite accurate and may be utilized periodically as a visual operating check of transmitter modulation during a program (reference 53).

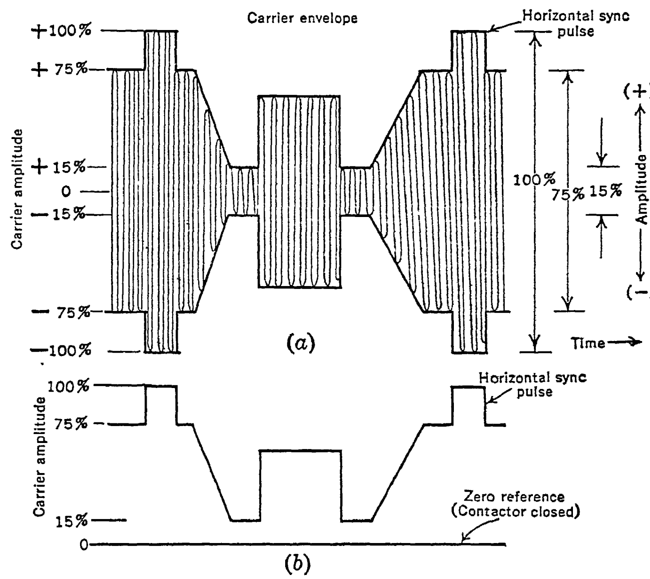


Fig. 24. Video Modulation Measurement. (a) Oscilloscope presentation of video modulated RF envelope (HF sweep). (b) Oscilloscope presentation of output of linear detector with shorting contactor (HF sweep).

MEASUREMENT OF R-F OUTPUT POWER. Television transmitters are rated in peak power output, i.e., the power output level attained during the synchronizing pulse portion of the transmitted signals.

The following methods of measuring power output assume that the transmitter power output can be held at the operating peak level. For transmitters that cannot be held at peak output for measurement purposes, other methods have to be used or a reliable correction factor must be applied.

The transmitter output power may be determined by measuring the power delivered to a water-cooled resistance load with circuit adjustments capable of transmitting a good picture. The rate of water flow and the temperature rise of the water stream on passing over the resistor must be accurately measured. Then the power delivered is given by

$$P = 263TF \tag{7}$$

where P = power delivered to load in watts.
T = temperature change of water in degrees centigrade.
F = water flow in gallons per minute.

It is often desirable to know the power delivered to the actual operating load. Since most practical transmission line installations are not perfectly matched to the radiator, a finite reflection occurs on the line. It is, therefore, necessary to determine the average value of transmission line voltage. One satisfactory method utilizes a slotted section of transmission line and a calibrated vacuum-tube voltmeter for determining the maximum and minimum values of the transmission line voltage.

The power delivered to the transmission line load, neglecting line attenuation, may be calculated by means of the following formula, which is accurate to better than 1 per cent if the voltage standing wave ratio is between 0.8 and unity.

$$P_0 = \frac{(E_{\text{max}})(E_{\text{min}})}{Z_0} \tag{8}$$

where P_0 = power delivered to load in watts.
E_{max} = rms value of voltage maximum in volts.
E_{min} = rms value of voltage minimum in volts.
Z_0 = characteristic impedance of transmission line in ohms.

Alternatively, the output power may be determined by permanently locating two calibrated vacuum-tube voltmeters precisely one-quarter wavelength (electrical) apart and substituting the voltage indicated by these meters for E_{max} and E_{min} in the above formula.

TRANSMISSION LINE. The transmission line system between the transmitter and the antenna must be well matched over the band of frequencies which includes the carrier and the sideband frequencies of appreciable magnitude. Multiple images may appear in the radiated signal of a poorly matched transmission system. Satisfactory results are usually produced when a standing wave ratio between 0.9 and 1.0 over the required band of frequencies exists.

ANTENNAS. Commercial television broadcasting antennas are required to be horizontally polarized. The directivity and radiating efficiency of the antenna should be substantially independent of frequency over the desired transmission band. The input impedance of the antenna must be substantially independent of frequency and must match the transmission line well enough to avoid the development and transmission of multiple images (references 54 and 55). Appreciable effective power gain may be obtained by compressing the radiated energy in the vertical plane.

PERFORMANCE MEASUREMENTS. The response of the visual transmitter from input to radio monitor may be measured using sinusoidal modulation. The modulating frequency should be varied incrementally over the required band while the relative response as a function of frequency is measured on a cathode-ray oscilloscope or vacuum-tube voltmeter. Alternatively, the sinusoidal modulating signal may be injected in the studio equipment before the blanking signals are added, and the relative response at the transmitter input and at the radio monitor output may be measured on a cathode-ray oscilloscope. This method is applicable to transmitting systems which use d-c restoration circuits, and the results are representative of what may be expected under actual operating conditions.

Means of observing and recording the transient response of a television transmission system are desirable. If a 100-kc square wave, having a rise time which is short compared to the rise time expected from the circuit under test, is applied to the transmitter the transient response may be observed on a cathode-ray oscilloscope connected to a r-f monitor. Some square-wave generators provide not only a 100-kc square-wave output but also a synchronous 100-kc sinusoidal output and a synchronous 10- or 20-Mc output. The 100-kc sinusoidal output may be advantageously used for horizontal deflection of the measuring oscilloscope. The 20-Mc output may be used to modulate the cathode-ray beam in amplitude. When the cathode-ray-tube bias is properly adjusted only the positive peaks of the 20-Mc modulation are visible, thus providing an accurate time base which may be used to measure the change in amplitude of the signal for each accurate time interval. The realized rise time, as well as the magnitude of overshoots or oscillatory transients, may thus be accurately determined. Such performance measurements can generally be correlated directly with the appearance of the reproduced television image.

TELEVISION RECEIVERS

By W. F. Bailey and R. J. Brunn

Receivers for television signals in accordance with the present-day standards of the Federal Communications Commission are of the superheterodyne type and receive both the picture and sound transmissions. A block diagram of a typical television receiver is shown in Fig. 1.

The picture and sound carrier signals are received by a single antenna and are amplified in a single channel. The selectivity of this channel protects against image signal and cross-modulation interference. Frequency conversion and some amplification at the intermediate frequencies are also accomplished in a single channel. Then the picture and sound signals are separated and each is amplified sufficiently for final detection. The selectivity in each channel must be adequate to keep the sound and picture signals from interfering with each other, and also to attenuate adjacent signals to a non-interfering level. An f-m detector and an audio amplifier complete the sound channel.

In the picture channel, following the second detector, amplification occurs at video frequency. Either direct coupling, or d-c restoration, or a combination, is used to maintain the voltage corresponding to black constant at the picture tube. In first-grade receivers, both automatic gain control and noise limiting are provided in the picture channel.

Synchronizing signals are extracted from the complete picture signal and are separated for the respective scanning oscillators.

Scanning generators produce sawtooth waves at line and field frequency of either current or voltage, depending on the type of picture tube. Magnetic deflection is commonly used for best resolution since there is less defocusing than there is with electric deflection.

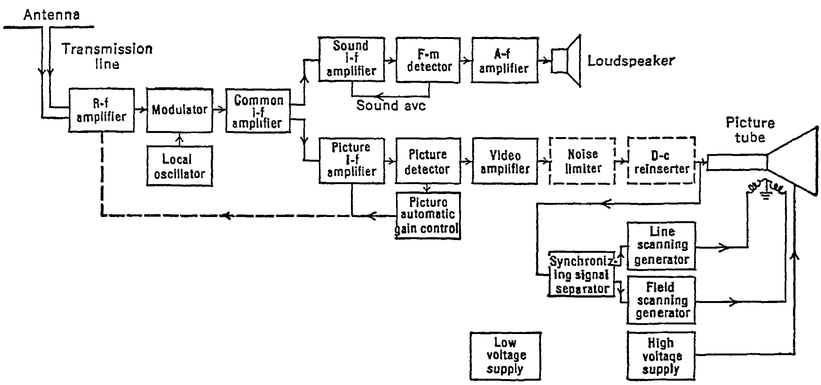


FIG. 1. Block Diagram of a Television Receiver

The d-c accelerating potential for the picture tube is obtained by rectifying the power-frequency wave, an r-f sine wave, or the voltage impulse during the line retrace in a magnetic scanning system.

23. ANTENNAS

The usual antenna for reception is a half-wave dipole in which the radiator diameter is from 0.2 to 2.5 per cent of the antenna length, to improve performance over the frequency range. The folded dipole antenna (reference 55) is also used with a 300-ohm transmission line. Some use has been made of a reflector to improve the directivity, but, with a simple array, not much directive gain can be obtained over the frequency range.

Usually, a balanced line (reference 57) of 75 to 300 ohms impedance is used to transmit the signal from the antenna to the receiver.

24. R-F CIRCUITS

REQUIREMENTS. The r-f circuits of a television receiver couple the signal from the antenna transmission line to the modulator. The following factors must be considered in the r-f circuit design: (1) r-f gain; (2) band width; (3) selectivity; (4) coupling to transmission line; (5) station selection; (6) oscillator radiation; (7) noise factor.

R-F AMPLIFIER. Normally, it is not necessary to provide amplification at radio frequency because it is easier to obtain the necessary amplification at intermediate frequency. However, an r-f stage is helpful in reducing oscillator radiation (reference 58) since there may be an attenuation of 10 to 50 times for signal propagation in the backward direction through the r-f stage.

Most modulators have considerable noise (references 59 and 60). The inherent noise of the receiver may sometimes be reduced by use of an r-f stage (reference 61). A triode r-f amplifier will reduce the noise to the minimum. Generally, a pentode r-f amplifier will provide no improvement. In order to eliminate the need for neutralization, the triode is generally used in a grounded-grid circuit, and a tube having a low plate-to-cathode capacitance is chosen.

ANTENNA COUPLING. With a low-loss transmission line, it is desirable that the receiver input circuit match the line with a low standing wave ratio to eliminate ghosts in the picture caused by multiple traversals of the transmission line (reference 62). The antenna cannot be expected to terminate the line with a standing wave ratio lower than about 10 db in some of the channels. Thus the reflection at the receiver must be kept low. With 2-db attenuation in the line in one traversal, and an antenna termination with a 10-db standing wave ratio, it is necessary that the receiver terminate the transmission line with a standing wave ratio of 0.6 db so that the signal-to-ghost ratio in the receiver be 40 db.

A resistive element must be present in the receiver to achieve termination of the transmission line with a low standing wave ratio. The input conductance of the first tube, the

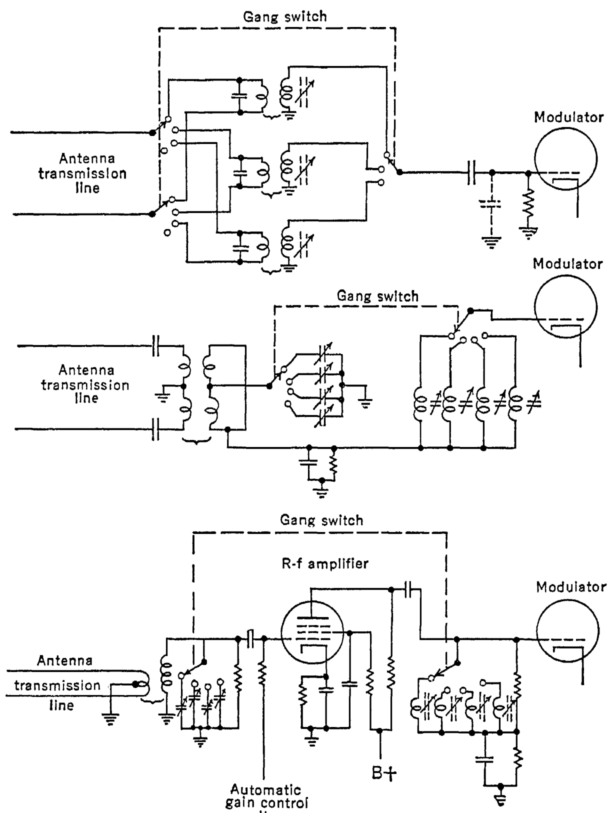


FIG. 2. Typical Antenna to Modulator Coupling Circuits

inherent losses in the reactive circuit elements, and, in some cases, a resistor added for the purpose, constitute the source of loading for the input circuit. The added resistor to produce the required loading increases the noise factor above the minimum (reference 63).

To realize the benefit of the balanced line in minimizing extraneous pick-up, a coupling circuit must be used between the transmission line and the first tube which has good transmission for balanced signals and high attenuation for unbalanced signals. This will reduce interference picked up by the transmission line acting as a single-wire antenna.

Transformers in which the electrostatic coupling is minimized are used to couple to the balanced line. The secondary is usually unbalanced to deliver the signal to a single grid or cathode. The transformer may be coupled directly into a cathode with a good impedance match, but no selectivity.

SELECTIVITY REQUIREMENTS. Because the picture and sound carriers are at opposite ends of a television channel, the r-f circuits should have substantially uniform transmission over a band width of about 5 Mc. This is required because the sound carrier, the main sideband, and the vestigial sideband should be amplified uniformly.

Sufficient selectivity should be provided to give an image ratio of at least 40 db. This provides protection against image signals which are as strong as the desired signal. This order of selectivity requires a minimum of two tuned circuits.

STATION SELECTION. Station selection may be accomplished by switching, with fixed or movable coils, or by continuously adjustable inductors, using tuning cores, or by varying the length of the wire to change the inductance.

The requirements for tuning are: (1) station selection by a single control; (2) reliable long-life, noise-free operation; (3) provision for a number of channels lying between 4 and 12; (4) resettability.

Some typical r-f circuits between the antenna transmission line and the modulator are shown in Fig. 2 (references 64 and 65).

25. MODULATOR AND LOCAL OSCILLATOR

MODULATOR. The tubes commonly used for the modulator are the triode and pentode types. Multigrid converters are rarely used because of their high noise (references 59 and 60).

Because it has the lowest internal noise, the triode modulator is used when the best noise factor is desired. However, the triode presents design problems, since both the input and the output impedances are functions of the oscillator excitation, and the grid-to-plate capacitance makes the input and output circuits interdependent.

The pentode modulator does not produce as low a noise factor because of the partition noise. It is less difficult to use in the receiver, since there is negligible coupling between the input and output circuits. Also, the output impedance of a pentode modulator is normally so high that variations in it, caused by changes of oscillator excitation, have no effect on the performance of the i-f amplifier.

A high transconductance tube is used to maintain high conversion gain and low noise. It is usual to bias the modulator by drawing grid current on the local oscillator signal.

LOCAL OSCILLATOR. The local oscillator is usually a triode used with either capacitive or inductive feedback. Capacitive feedback offers the advantages that the inherent capacitances of the tube may be used directly in the oscillator circuit to produce feedback, and that the tuning coil has no tap.

LOCAL OSCILLATOR DRIFT. Both the picture and sound quality suffer if the oscillator varies from the correct frequency either by oscillator drift or poor resettability of the tuning device. Frequency stability is of prime importance. A frequency shift of ± 150 kc is about the upper limit that can be tolerated by the picture. This drift will produce approximately ± 2-db variation in the amplitude of low-frequency video components relative to high-frequency components. The cost of the sound channel is increased with high shifts. The sound-channel band width and the linear portion of the detector characteristic must be adequate to accommodate the drift. Otherwise the sound i-f may lie on the side of the i-f amplifier transmission characteristic. This produces amplitude modulation of the f-m signal. Further, the performance of the frequency detector may suffer, as the signal may be near one of the peaks and will be on a non-linear part of the detector characteristic.

Receivers have frequently employed an oscillator tuning adjustment so that the resettability errors of the station-selecting device and the oscillator drift can be corrected by the user.

OSCILLATOR INJECTION. The oscillator signal is injected into the modulator grid circuit by either magnetic or capacitive coupling. In many cases, the stray capacitance of the station-selector circuit wiring is sufficient.

26. PICTURE I-F AMPLIFIER

The picture i-f amplifier provides most of the amplification required and also controls the frequency band of the signal in the picture channel. The pass band varies in width from about 2 Mc in a low-definition receiver to about 4 Mc in a high-definition receiver. The i-f amplifier attenuates signals on the adjacent channels so that these signals do not interfere with the picture.

FREQUENCY. The choice of the frequency band for i-f amplification is governed largely by interference from direct i-f pick-up and image signals. The local oscillator is usually located on the high-frequency side of the signal as this simplifies the image interference problem, and thus the picture intermediate frequency is higher than the sound intermediate frequency. A choice of at least 20 Mc for the sound intermediate frequency eliminates other television stations as images, but the f-m broadcast stations are then in the image-signal range. A choice of 29 Mc or higher eliminates both television and f-m broadcast stations as image signals.

The approximate frequencies of the i-f band are chosen with regard to image signals, and the exact frequencies are chosen to eliminate serious interference by direct i-f pick-up. Of the sources of strong signals in the bands cited above, amateur frequencies are worst, since the transmitter may be close to the receiver.

Most current designs utilize a frequency of about 26 Mc for the picture i-f carrier.

VESTIGIAL SIDEBAND REQUIREMENTS. For band-width conservation, picture signals are transmitted by a vestigial sideband system (reference 66. See also article 9). In this transmission system, it is necessary to attenuate the carrier frequency 6 db relative to the transmission of the major sideband, and to adjust the transmission of both sidebands adjacent to the carrier so that a uniform output-frequency spectrum results when a uniform input spectrum is applied to the system. This adjustment is made in the receiver.

In most receivers, the carrier and low video-frequency transmission is equalized prior to detection as shown in terms of the r-f signal by *B* of Fig. 20, Article 9. Over a frequency band of about 1.5 Mc, the transmission drops from full value to 10 per cent or less, with the cutoff characteristic so chosen that the carrier is transmitted at 50 per cent of full transmission.

It is desirable that this cutoff characteristic be as gradual as the standards allow. This reduces the distortion produced by the quadrature component (references 51, 22, and 23) and the non-linear phase characteristic associated with the amplitude cutoff.

ATTENUATION CHARACTERISTIC. A typical picture i-f amplifier transmission characteristic is shown in Fig. 3. The attenuation of the desired sound carrier, which is

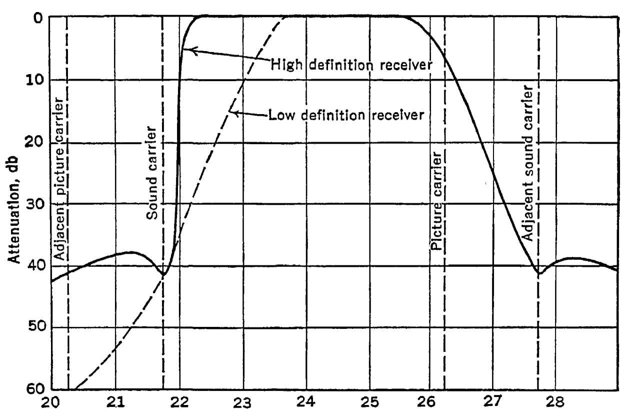


FIG. 3. Typical Receiver Picture I-f Response

normally obtained by the use of traps in the i-f coupling impedances, should not be too steepsided, since a high cutoff slope will convert the sound-frequency modulation to amplitude modulation, which may show in the picture. The sound attenuation should be about 300 kc wide, 3 db above the minimum, so that frequency drift of the local oscillator will not cause sound interference in the picture.

It is desirable that about 40 to 50 db total attenuation be provided for the sound carrier. This may be produced entirely by i-f selectivity, or it may be obtained partly in the i-f amplifier and partly by attenuation at 4.5 Mc in the video amplifier.

Present station-assignment practice is such that adjacent channels will not be allocated in any region. The overlapping service areas of any two regions whose allocations occupy successive channels are small. Therefore, it appears reasonable to provide attenuations of about 35 db minimum relative to the desired picture carrier for the adjacent channel carriers.

It is usually necessary to use traps to secure the attenuation at the sound carrier of the lower-frequency adjacent channel.

COUPLING NETWORKS. Coupling networks used in the i-f amplifier are of several forms: double-tuned transformers (reference 67), filter-type networks, and stagger-tuned resonant circuits are commonly used. See Section 7 for more information.

Various methods are used to incorporate traps in the i-f amplifiers. Traps may be part of the coupling impedance, or they may be used to reduce the effective transconductance of the amplifier tube.

Figure 4 shows several circuits in which traps are employed to reduce the transmission in a desired frequency range. The transmission characteristics are also shown. Figure 4*A* shows a single stage in which a stagger-tuned single resonant circuit is used. The grid leak of the following tube is chosen to provide the required Q. An inductively coupled

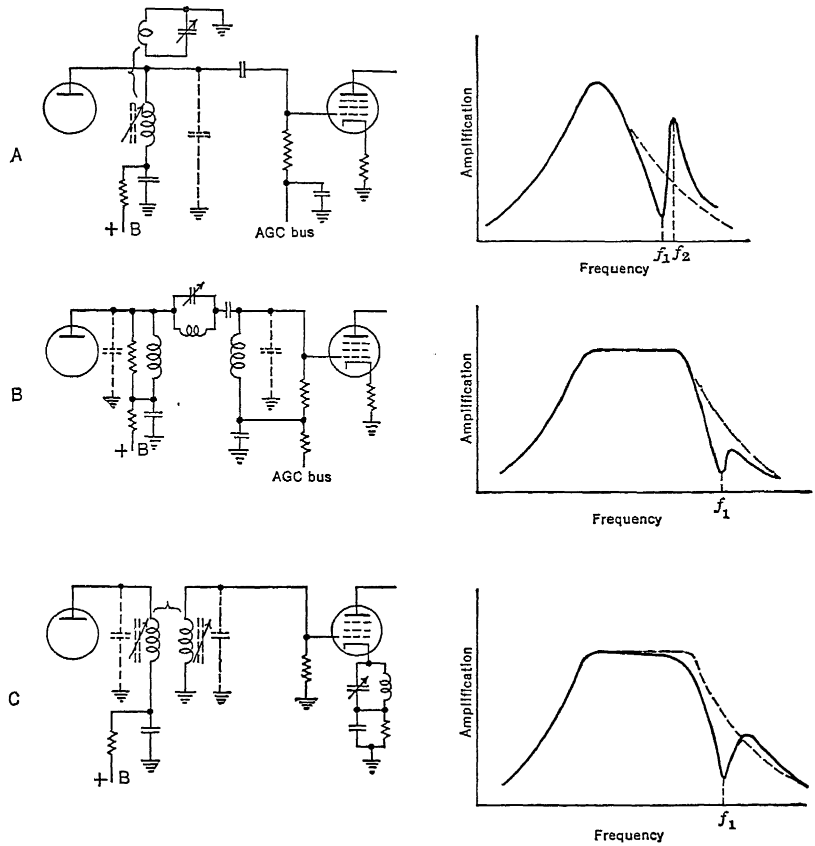


FIG. 4. Typical I-f Coupling Networks

trap is used to secure attenuation at frequency f_1. An undesirable feature of the inductively coupled trap shown is the spurious response at frequency f_2. The magnitude of this spurious response is proportional to the Q and coupling of the trap, and its maximum can be substantially the same as the main response.

Figure 4*B* shows a single stage in which the coupling impedance is a section of an m-derived filter. The response may be made uniform over the desired band with attenuation at a specified frequency f_1.

Figure 4*C* shows a single stage in which the coupling impedance is two coupled circuits. To secure maximum gain, the damping is concentrated on one circuit only. Attenuation at frequency f_1 may be produced by a parallel resonant trap in the second tube cathode circuit. Such a trap usually affects the input impedance of the tube because of feedback to the grid circuit. It is undesirable to use cathode traps with tubes in which the suppressor grid is connected to the cathode. As the cathode has considerable impedance to ground, the suppressor-to-anode capacitance may couple sufficient signal from the output circuit back to the input circuit to cause instability.

As the trap attenuation is a function of both the transconductance of the tube and the impedance of the trap, cathode traps are normally employed in fixed-gain stages.

GAIN CONTROL. Gain control is usually accomplished by varying the bias of one or more i-f amplifier tubes. It is generally necessary, because of the use of high-transconductance tubes, and circuits in which the tube capacitance is a large part of the total capacitance, to stabilize the input capacitance of the gain-controlled stages by an unbypassed cathode resistor (reference 68).

27. PICTURE CHANNEL SECOND DETECTOR

The second detector in the picture channel is usually of the diode type. The video-frequency output-signal load impedance is determined by the shunt capacitance of the diode and the following stage and by the band width to be transmitted. If the impedance band-width product is too low when the total capacitance is lumped as a single capacitor to ground, then video-filter technique may be used. The total capacitance is then broken up into several smaller units, allowing the impedance band-width product to be increased. See Section 7.

DIODE LOAD. If the simple load circuit as shown in Fig. 5*A* is used, the rise time for the video output signal for outward modulation of the carrier will generally be shorter

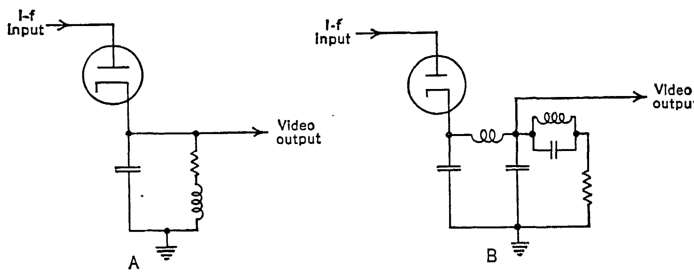


FIG. 5. Picture Detector Circuits

than the fall time for inward modulation of the carrier. The diode and generator resistance shunt the diode load time constant for outward modulation but not for inward modulation.

If the diode load circuit comprises a filter network as shown in Fig. 5*B*, the rise and fall times of the video output signal are more nearly equal because they are determined by the rate at which energy propagates through the filter network. The input should be at the open end of the filter, since, with this connection, there is minimum reflection in the filter which would affect the current supplied by the relatively low effective impedance of the diode and i-f output circuit.

There may be variations in the charging time with the video modulating frequency, since the output impedance of the i-f circuit driving the diode is not, in general, constant. This effect is usually not serious.

TUBE CONSIDERATIONS. The diode load impedance is low, ranging from about 2000 to 8000 ohms. The output voltage generally ranges from about 1 to 5 volts, and this results in high peak currents. To minimize the signal loss in the diode, it is thus desirable to use a high-perveance tube with low interelectrode capacitance.

I-F HARMONIC INTERFERENCE. The use of a four-terminal diode load impedance like that of Fig. 5*B* generally attenuates the ripple frequency component sufficiently so that it, or its harmonics, do not produce spurious patterns in the picture. These i-f harmonics can be troublesome on channels where they fall in the r-f picture frequency band, if they are fed back to the r-f section of the receiver with sufficient level, as a beat-frequency signal lying within the video-frequency passband of the receiver will then be produced. A beat takes the form of alternate dark and light bands in the picture. Since the beat signal is not related harmonically to the scanning rate, the bands continually move about the picture. With full-wave rectification the fundamental ripple frequency superposed on the video signal is twice the intermediate frequency and it is reduced in amplitude, which simplifies the filtering.

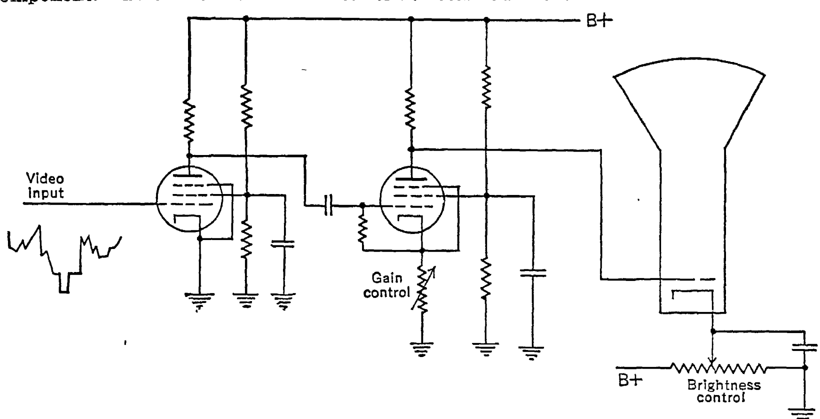
OUTPUT POLARITY. The diode detector may be arranged to produce a video output signal of either polarity. For the negatively modulated signals standardized in the United States, the detector circuits of Fig. 5 will deliver video output signals of negative polarity; that is, the synchronizing signals will be the most positive part of the video-frequency output.

28. VIDEO AMPLIFIERS AND DISPLAY

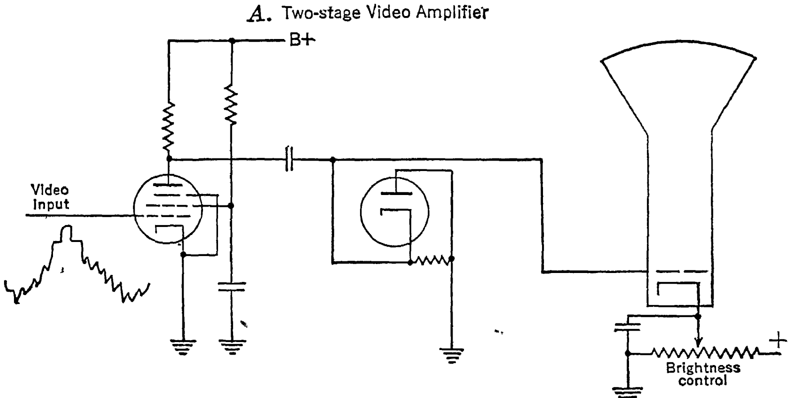
The video amplifier, in a television receiver, raises the level of the picture-detector output signal to a satisfactory value for the picture tube. The input level is commonly about 1 to 5 volts peak-to-peak, and the output level ranges from about 20 to about 100 volts peak-to-peak. It must be remembered that the signal range from black level to the synchronizing-signal peaks does not contain picture information. Thus the video amplifier must handle a complete signal about 40 per cent larger than the black to white signal. One or two stages of video-frequency amplification generally suffice in the usual television receiver.

FREQUENCY REQUIREMENTS. The video amplifier transmits a wide band of frequencies, one cutoff being at some low frequency, which may be direct current. The other cutoff is at some high frequency, usually lying between about 2 and 4.5 Mc, depending on the desired resolution.

Direct coupling is difficult to use in a multistage amplifier because of the problem of obtaining proper electrode potentials. It is simpler to design amplifiers whose lower cutoff lies in the range of about 10 to 10,000 cycles per second. Effective transmission of the d-c component of the signal may be accomplished by means of a locally generated d-c component. This involves a d-c reinserter as described in article 7.



A. Two-stage Video Amplifier



B. Video Amplifier with D-c Reinserter

FIG. 6. Typical Video Amplifiers

In Fig. 6*A* there is shown a two-stage video amplifier. D-c reinsertion is provided by the high grid-cathode conductance for positive grid potentials of the second amplifier tube. The picture tube is direct coupled to the output of the second video amplifier tube.

A gain control which operates by varying the amount of negative regeneration for alternating current is provided in the cathode circuit of the second stage. A change in gain of about 6 to 1 may be obtained with uniform frequency response with a control of this type. The rheostat should be non-inductive, and its range is limited by shunt capacitance which by-passes it for high video frequencies.

A potentiometer is sometimes used as an alternative gain control, in which the signal on the arm of the control is supplied to the amplifier grid. In this case, shunt capacitance to ground from the amplifier grid may vary the frequency response with gain-control setting.

Figure 6*B* shows a circuit in which the signal is a-c coupled from the video amplifier to the picture tube. A d-c reinserter is provided to stabilize the potential of synchronizing signal peaks on the picture tube grid.

For d-c amplifiers or those using d-c reinserters on the input circuits, it is essential that the amplifier stage have the same gain for direct current as for other frequencies within its passband. This requires that the cathode, screen, and anode supply potentials be stabilized against variation with varying direct current flowing in the amplifier tube. Failure to meet this requirement means that the instantaneous brightness of any part of the picture will be dependent upon the average brightness.

If the band width to be transmitted is not extreme, and the shunt capacitance not high, a simple two-terminal constant-*k* type of network is often used for a coupling impedance. This type may be designed for quite uniform amplitude and phase characteristics. When higher impedance is desired, or the circuit is required to work with high total shunt capacitance, it is common to use four-terminal networks. This type allows the shunt capacitance to be broken into smaller lumps, thus giving a higher impedance-band width product. For maximum exploitation of the band width, the circuit is designed as a filter (reference 69). Such a filter, while it provides a maximum of uniform amplitude passband, has a fairly sharp cutoff characteristic and a non-linear phase curve, both of which may produce objectionable distortion in the form of echoes (reference 9) of the original signal. In general, better performance is obtained with a network in which the amplitude characteristic falls gradually with increasing frequency, as this reduces both the phase and amplitude distortion. Section 7 gives more specific information regarding the design of coupling impedances.

The video-frequency coupling network is generally designed to have uniform impedance if it is of the two-terminal type, or uniform transfer impedance if it is of the four-terminal type. Such designs give uniform gain if driven by high-impedance sources but do not produce uniform gain if the driving source impedance approximates that of the network. For the two-terminal type this is true because the network impedance is complex and has a variable phase angle over the transmitted band. For the four-terminal type this is true because, for uniform transfer impedance, the input impedance at the driving point is either uniform in magnitude but complex with a variable phase angle over the transmitted band, or non-uniform in both magnitude of impedance and phase over the transmitted band.

Normally, video amplifiers use the grid-cathode circuit for input, and the plate-cathode circuit for output, and thus the signal polarity is reversed in going through a stage. In the design of a receiver, the picture detector must be so poled that the desired polarity is obtained at the picture tube grid.

PICTURE TUBE. The present-day picture tubes are of the cathode-ray type. The electron beam is focused by an electron gun which may utilize electric fields only or a combination of electric and magnetic fields. Deflection of the cathode-ray beam is produced by either electric or magnetic fields. In the present state of the art, magnetic focus and deflection appear to give the best performance in regard to: (1) spot size; (2) high current in the beam; (3) uniformity of focus over the raster.

For direct-view receivers, the final anode voltages range from about 3 kv to about 15 kv. For projection-type receivers, the final anode voltage in current designs is about 30 kv.

The phosphor produces a white light which may vary in shade from slightly bluish or greenish to yellowish.

It is usual to provide a bias control to adjust the average brightness of the picture. Examples of this are shown in Fig. 6. Sufficient bleeder current flows through the bias control so the picture-tube current does not vary the bias appreciably, and a by-pass is provided for high-frequency currents.

Section 15 contains more detailed information on picture tubes.

PICTURE GAIN CONTROL. Automatic gain control for the picture channel is desirable in television receivers, as it minimizes readjustment of the controls when switching from one channel to another. Radiated signals conforming to the FCC standards include the d-c component. Thus, the average carrier level is dependent upon the picture content as explained in articles 7 and 9. It is necessary that the automatic-gain-control circuit respond to a part of the signal which is independent of the transmitted picture. With a negative polarity signal, as prescribed by the FCC standards, it is most convenient to develop the automatic-gain-control voltage from the synchronizing signal peaks. This requires that the automatic-gain-control rectifier load circuit have a time constant of not less than several lines duration, so that the picture content cannot affect the automatic-gain-control voltage. A separate rectifier operated at the same level as the picture detector may be used as the source of automatic-gain-control voltage. Better performance may be obtained by amplifying the rectifier output voltage with a d-c amplifier. Figure 7 shows

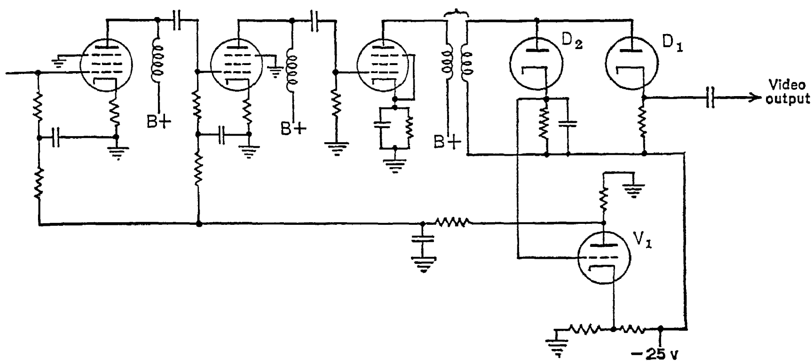


FIG. 7. Picture Automatic Gain Control Circuit

such an arrangement. In this circuit, D_1 is the picture detector; D_2 is a separate diode with a high-impedance load having a time constant of about 200 μsec. The output of D_2 is direct coupled to a triode V_1, the cathode of which returns to a negative potential, in this case, 25 volts. As the signal level increases, the anode of V_1 falls in potential, providing an amplified voltage which is suitable for an automatic-gain-control bias. An alternative of this circuit may be used, in which the additional amplification occurs prior to the automatic-gain-control rectifier. This amplification may take place at intermediate frequency or video frequency.

29. NOISE LIMITERS

Noise limiters are sometimes used to reduce the effects of impulse noise interference upon the picture-tube signal and upon the synchronizing performance. It is desirable to limit the impulse noise to a level no greater than that of the synchronizing signal peaks so that the operating bias of the video amplifier or the synchronizing signal separator is not changed. In circuits where direct coupling is used, impulse noise generally does not greatly affect the operating characteristics. A-c coupled circuits are usually affected considerably by noise.

Diodes connected in shunt or series in the video amplifier have been used as impulse noise limiters An example of a shunt-connected diode limiter is shown in Fig. 8. As the

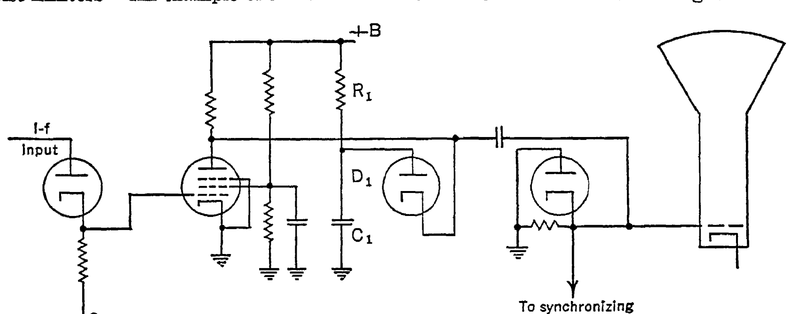


FIG. 8. Impulse Noise Limiter

video amplifier is direct coupled to the picture detector, its operating conditions are not seriously affected by the noise. The limiter diode D_1 is connected in its anode circuit. The d-c reinserter for the picture tube, and the synchronizing signal separator, are actuated by the signal following the limiter and thus operate with increased reliability. The limiter shown adjusts itself to the signal level and normally limits the peak of each synchronizing pulse slightly. If the noise has a high duty cycle, this type of limiter will fail, as the noise will then begin to bias the diode D_1 off, since resistor R_1 will not be able to remove the added charge from C_1 quickly enough. By stabilizing the potential of the anode of D_1 with a bleeder, the limiter will handle noise of high duty cycle but will not adjust itself to the signal level.

30. SOUND AMPLIFIERS

I-F CIRCUITS. The sound i-f amplifier of a television receiver must provide adequate gain with proper selectivity characteristics for the f-m sound signals which accompany the picture. The design features of television receiver sound i-f amplifiers depend largely on the receiver type, whether broadcast a-m or f-m services are to share the channel, and the amount of gain provided by the circuits which precede the point of sound i-f take-off.

Gain Requirements. The sound circuits should be capable of providing a comfortable audio output with 30 per cent modulated sound carriers from 6 to 10 db weaker than the threshold picture level. Where manual or automatic picture-gain-control circuits can reduce the amplification in the overall sound channel by operating on tubes ahead of the sound take-off point, an additional margin of sound gain is required.

With present techniques, the threshold picture level is of the order of 50 μv. An overall sound sensitivity of about 10 μv would therefore seem suitable for television, although an additional 20 db might be desirable for broadcast frequency modulation.

When switched for television, receiver front-end circuits seldom develop more gain than is lost by the modulator in converting to the intermediate frequencies. The sound i-f sensitivity on the modulator grid is therefore about the same as the overall sound sensitivity. The amount of sound-channel gain that may be provided between the modulator grid and the point of sound take-off varies widely with receiver designs. Some picture i-f amplifiers can provide between 40 and 50 db of sound-channel gain in the modulator and first one or two common stages. The most serious objection to this arrangement is the conflict that results from manual or automatic picture gain control of these stages.

The output of the sound i-f usually feeds either a ratio-type f-m detector or a limiter working into a conventional f-m detector. The minimum i-f output required depends on the detector and/or limiter design and is usually in the range of 1 to 3 volts.

Selectance Characteristics. The sound i-f amplifier of a television receiver must be broad enough not only to pass the sidebands of the carrier with full 25-kc deviation but must also pass this signal when the local oscillator is detuned because of drift or inaccurate resettability of the tuning device. Minimum 6-db band widths between 200 and 400 Kc are usual.

For television service, the sound-channel selectors should provide at least 20-db attenuation at the picture carrier and 50 db or more against signals on adjacent channels. This is considerably less severe than the requirements for broadcast frequency modulation as outlined in Section 8. Where dual service is contemplated, the selectivity requirements should be based on the broadcast frequency modulation, and it is then necessary to keep the local oscillator frequency drift within the band width provided.

Sound Take-off Methods. Television receivers which pass 3.5- to 4-Mc video band width usually require traps to provide sufficient attenuation of the sound intermediate frequency in the picture channel. Such traps usually build up a sound i-f voltage or current, and they may be coupled either directly or through additional circuit elements into the grid of the first sound i-f amplifier. This method is applicable to simple coupled traps, to cathode traps, and to coil arrangements in stages coupled by filter circuits.

Receivers passing less video band width may not require sound traps. The sound take-off may then be from the secondary of a transformer whose primary is connected in series with a picture i-f transformer; or the sound and picture i-f amplifier grids may be operated in parallel.

Amplifier Design. There is usually negligible selectivity for the sound intermediate frequency in the common picture and sound circuits. Some selectivity may be designed into the take-off circuits. When its gain and selectivity requirements have been established, the sound i-f amplifier can be designed by the techniques described in Section 7. An adequate number of single- or double-tuned circuits can give acceptable performance provided that 20 to 30 $\mu\mu$f over stray capacitance is added to each circuit. The alignment procedure will be simplified if the circuits are under-optimum coupled.

The mistuning of gain-controlled stages should be minimized either by means of unbypassed cathode resistors or by tapping down the grid.

SOUND DETECTOR AND AUDIO AMPLIFIER. The design of the sound detector and the audio amplifier for a television receiver follows broadcast f-m practice as discussed in Section 8. Since 100 per cent modulation on a television sound carrier corresponds to 25 instead of 75 Kc deviation as in broadcast frequency modulation, only one-third the output voltage is obtained from equivalent f-m detectors. In addition, the output performance of f-m detectors is usually degraded when the carrier frequency is increased, as in television sound i-f amplifiers. The gain deficiency can sometimes be made up by employing high-gain audio amplifier circuits, although this is usually undesirable, as hum-pickup difficulties are inevitable.

An undistorted electrical output of 1 watt is probably adequate for many home television receivers, as the audience is close to the receiver. Television receivers incorporating broadcast f-m or a-m are usually capable of providing greater power output.

Since the current drawn by an output audio amplifier varies with the signal, and since this current may represent a sizable fraction of the total B current drain, special consideration must be given when video circuits obtain power from the same B supply. Either adequate decoupling arrangements must be made or a constant current output amplifier circuit must be used.

31. SYNCHRONIZATION

Adequate synchronizing circuits are among the most important features that a television receiver must possess. The least expensive receiver must be capable of synchronizing on any signal of reasonable strength without readjustment of the speed controls. More expensive receivers may be expected to maintain synchronization on threshold weak signals in the presence of interference.

The procedure for effecting synchronization in the television receiver consists, first, of extracting the synchronizing pulses from the complete video wave. The line and field pulses are then usually separated from each other and the resulting signals are used to synchronize the respective scanning oscillators. See article 10 for a discussion of this.

SEPARATION OF SYNCHRONIZING PULSES. One method for extracting the synchronizing information is to provide a separate diode detector for this purpose. The diode load resistor is by-passed by a capacitor proportioned so that the d-c voltage developed across the load resistor cannot drop more than about 20 per cent during a line interval. The charging current in the capacitor is then a measure of signals in excess of 80 per cent of the peak amplitude of the carrier. A voltage proportional to the charging current may be obtained across a small resistor in series with the capacitor.

This type of separator exaggerates amplitude modulation of the synchronizing signal pulses which may be present in the complete signal. Additional amplification and limiting are usually required.

As full video band width is not required for the synchronizing pulses, the separate detector can be preceded by a high-gain narrow-band-width stage. If sufficient signal is developed, voltage across the entire diode load, which is proportional to peak carrier amplitude, may be used for automatic gain control as described in article 28.

Video Separation. Synchronizing signals can be separated from a composite picture signal by the use of limiter circuits operated in conjunction with suitable d-c stabilization of the wave applied to the limiter. The most frequently used limiter of this type employs a sharp cutoff tube, usually a pentode, with the signal applied to the grid with black positive (reference 70). The tube is usually operated with a grid leak and blocking condenser input circuit, and without bias. Grid current is, therefore, drawn on the tips of the synchronizing pulses. The cutoff characteristic of the tube and the amplitude of the applied video wave are correlated so that the grid swing due to the synchronizing pulses alone exceeds the cutoff.

The self-bias d-c restoring method described above results in poor performance in the presence of impulse-type noise unless preceded by a suitable noise limiter. A strong noise pulse reaching the grid draws current and depresses the grid wave until the blocking capacitor can discharge. Several synchronizing pulses can thus be lost.

The synchronizing information may be extracted from picture signals of either polarity by diode circuits, examples of which are shown in Fig. 9. The time constant of R_1 and C is proportioned so that the voltage across R_1 drops to about 80 per cent during a line interval. The diode then conducts only during the synchronizing pulse. A voltage corresponding to the diode's conduction is obtained across the proportionately smaller resistor R_2 in series with the diode.

The amplitude of the output synchronizing wave from diode separators of this type varies with picture content and usually requires amplifying and limiting for good synchronization. The interelectrode capacitance of the diode may couple high video-frequency component current into resistor R_2.

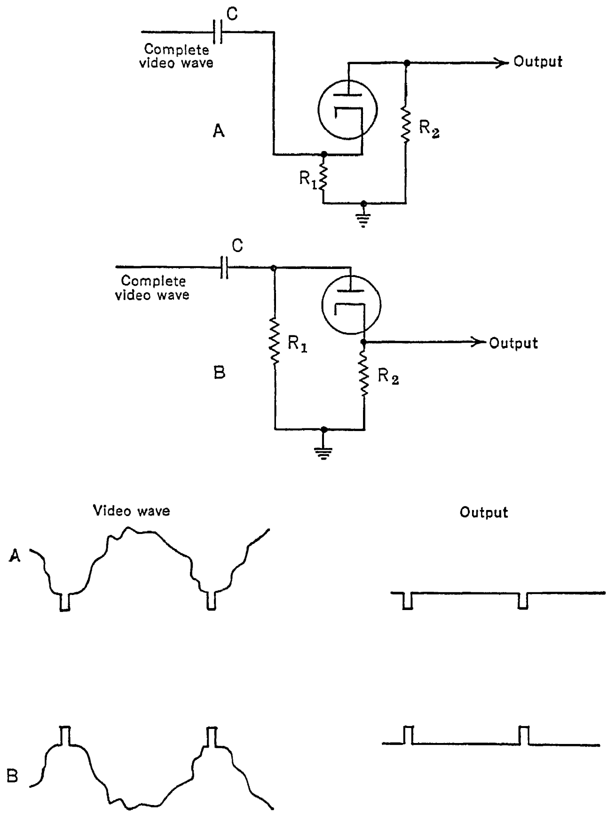


FIG. 9. Diode Synchronizing Separator

SEPARATION OF LINE SYNCHRONIZING PULSES. The line synchronizing pulses are usually separated from the complete synchronizing signal by differentiation. A typical differentiating circuit is shown in Fig. 13, p. 20-13. This operation produces a wave containing a series of narrow pulses coincident with the leading edges of the equalizing pulses, the line synchronizing pulses, and the broad field synchronizing pulses, as shown in Fig. 12, p. 20-12, to assure continuous operation of the line scan oscillator throughout the field retrace interval.

SEPARATION OF FIELD SYNCHRONIZING PULSES. Field synchronizing pulses can be separated from the complete synchronizing signal by integration. To preserve reasonable rise time of the output pulses and still eliminate the line pulses, a multistage integrator is sometimes used. An example is shown in Fig. 13, p. 20-13.

OSCILLATOR SYNCHRONIZATION. Triggering. When the triggering technique is used, the oscillators employed are usually types that free-run at slower than the correct speed and the synchronizing pulses are applied to initiate the retrace. The oscillators should be designed to be insensitive to triggering except towards the end of the trace so that the possible mistiming is limited to the interval between the oscillator's sensitivity to triggering and its self-retrace. Multivibrators, blocking oscillators, and thyratron oscillators are commonly used. Where oscillator voltage appears on the triggering terminal, buffers are usually required.

Good performance is obtained from triggered oscillator circuits only when the synchronizing waves are clean. Video components and other extraneous signals should be small. The effects of random noise can be minimized by restricting the passband into the synchronizing circuits. Impulse noise should neither greatly exceed in amplitude, nor cause a loss of, the synchronizing pulses after the disturbance.

Phase Control. Phase-controlled scanning circuits employ oscillators whose frequency can be controlled by a d-c voltage (reference 17). Scanning oscillators can usually be so controlled through the use of a d-c amplifier or a control tube. The control voltage is obtained by measuring the phase difference between the synchronizing signal and a signal from the scanning oscillator.

The advantage in phase-controlled synchronizing is that an extremely narrow passband can be employed in the coupling between the phase comparison circuit and the oscillator control point to make the oscillator insensitive to instantaneous aberrations of the synchronizing wave. Random noise, impulse noise, and, usually, small amounts of video can be tolerated. The use of a sufficiently narrow passband to achieve the desired degree of stability tends to result in a sluggish pull-in characteristic. Typical performance is to require a second or more to lock.

32. SCANNING

Conventional scanning circuits for television receivers usually employ scanning oscillators and output amplifiers. The oscillator output pulses are shaped as required and applied to the grids of output amplifiers (reference 71) to produce voltage or current waves of proper magnitude and shape.

To achieve economies, the functions of the oscillator, wave shaper, and output amplifier are sometimes integrated, as in the circuits shown in Figs. 9 and 11 of pp. 20-9 and 20-11.

SAWTOOTH WAVE-SHAPING CIRCUITS. The voltage wave required on the grid of scanning output amplifiers departs from being of sawtooth form only by what is required to correct for the deficiencies in the output circuit. The usual method in scanning generators is to integrate current pulses in a capacitor.

A typical sawtooth generator is shown in Fig. 10. The tube is normally cutoff. When a positive pulse is applied to its grid, the capacitor in its plate circuit is discharged. Following the pulse, the capacitor recharges to the supply voltage as shown at *A*. By using a time constant 5 or 10 times as long as the interval between pulses, a reasonably linear sawtooth wave may be obtained.

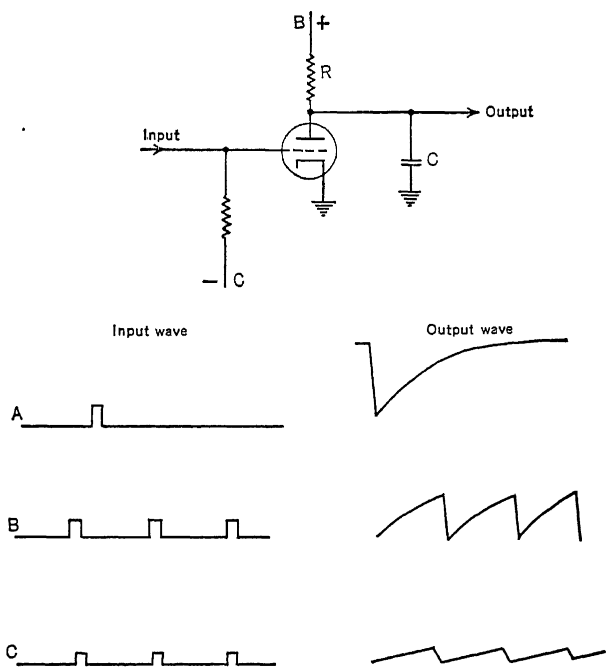


Fig. 10. Sawtooth Wave Generator

SCANNING OSCILLATORS. Blocking Oscillators (reference 72). The blocking oscillator has been the preferred scanning oscillator in television receivers. A blocking oscillator is shown in Fig. 11. When the transformer windings are connected so that the grid goes positive when the plate goes negative, this circuit will start oscillation and will generate a pulse. During the pulse, grid current flows and charges the capacitor C negatively, eventually terminating the pulse. The capacitor then discharges through the resistor R to initiate a new cycle. The free-running speed is controlled by the capacitor, the resistor, and the voltage B_1.

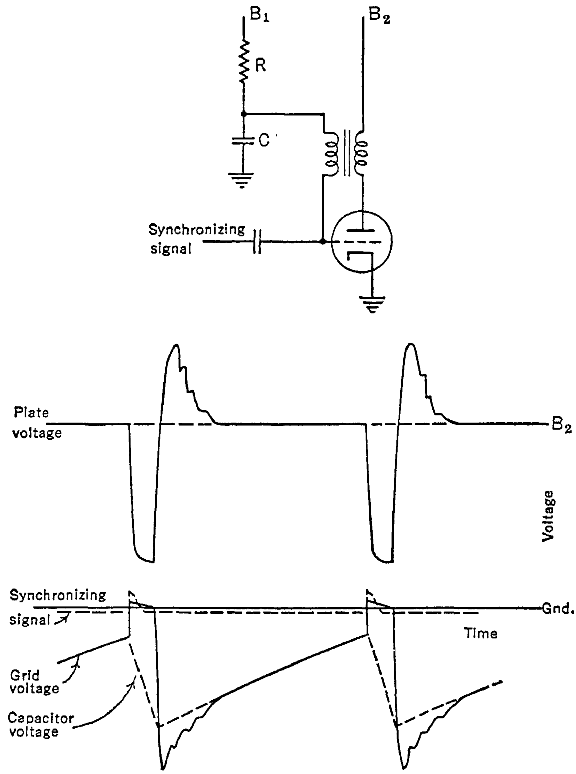


FIG. 11. Blocking Oscillator

The blocking oscillator may be synchronized by applying pulses to initiate conduction in advance of the capacitor discharge. For phase-controlled scanning circuits, the speed of the blocking oscillator can be regulated by controlling the voltage B_1.

Multivibrators. Multivibrators are frequently used in low-priced receivers for economy. These circuits are generally regarded as being less stable, unless considerably more than the minimum number of essential parts are employed.

A multivibrator circuit arrangement useful for television receivers is shown in Fig. 12. The operation of this multivibrator is shown by the wave forms.

Since the current in tube B consists of a recurrent pulse wave, a wave-shaping circuit as described in Fig. 10 may be placed in the plate circuit of this tube for generating a sawtooth voltage wave, as shown in Fig. 12.

Thyratron Oscillators. Thyratron tubes filled with the lighter inert gases can be used as television scanning oscillators. The tubes are connected to discharge a capacitor in the plate circuit. Speed is controlled by varying either the recharge time constant or by the cathode bias. Synchronizing pulses can be applied to the grid.

The advantage of the thyratron tube lies in its ability to pass peak currents of high amplitude, but time delay circuits may be required to prevent application of anode voltage before the cathode has reached proper operating temperature.

OUTPUT AMPLIFIERS FOR ELECTROSTATIC DEFLECTION. The resolving capabilities of most electrostatic receiver tube types can be realized only when the average of the voltages on the deflecting plates of a pair is maintained equal to the second anode voltage. This necessitates balanced deflection as well as balanced centering circuits. The scanning output circuits for electrostatic tubes usually, therefore, produce sawtooth waves of both polarities.

Typical output amplifier circuits employ two voltage-amplifier tubes connected to give opposite polarity outputs. A separate phase inverter is seldom used; the second tube is

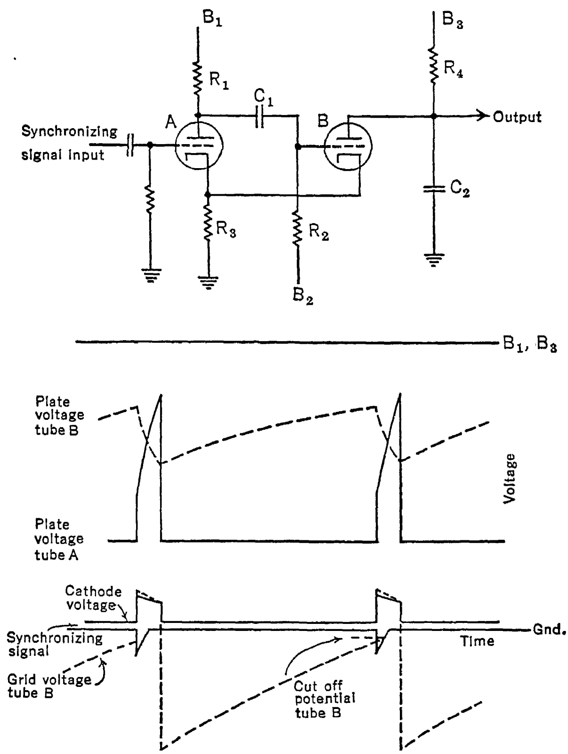


FIG. 12. Multivibrator Scanning Oscillator and Wave Shaper

usually driven by the first, either by the plate connection shown in Fig. 8, p. 20-9, or through the common cathode resistor. Where the total scanning voltage required exceeds the order of 600 volts, the B voltage required for the output amplifier tends to become quite high.

OUTPUT AMPLIFIERS, MAGNETIC. Since most magnetic scanning circuits operate by generating and dissipating energy during each scan, the total power and the circuits for line scanning are considerably different from those of field.

Linear magnetic deflection is accomplished by passing a sawtooth current through the windings of the deflection yoke. The internal resistance of the amplifier, non-linearity of the amplifier, and impedances in shunt or in series with the yoke usually require a wave form other than sawtooth at the amplifier grid. The grid wave is thus sometimes exponential and has a pulse component added by inserting a resistor in series with the capacitor in Fig. 10.

Line (reference 82). The usual receiver line scan circuit employs the idealized scanning cycle shown in Fig. 10, p. 20-10. Where high efficiency is required, the triode dissipating circuit shown in Fig. 11, p. 20-11, is used either with or without the "bootstrap" connection which reclaims some of the scanning energy.

Diode circuits are shown in Fig. 13. A high-perveance diode may be connected across the yoke as shown in *A*. Lower-perveance diodes, having adequate voltage rating, may be connected across the primary as shown in *B*.

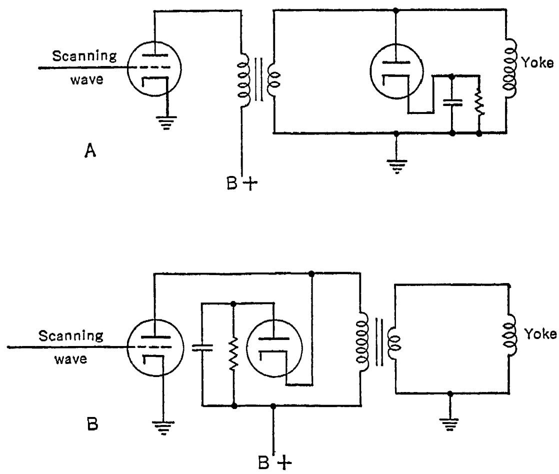


Fig. 13. Diode Damping Circuits

Field. The output load on the field amplifier is essentially resistive so that the controlled dissipation circuits used in line scanning are seldom employed. Occasionally, damping elements are placed across the yoke to remove transients after the retrace. A typical field scanning circuit is shown in Fig. 14.

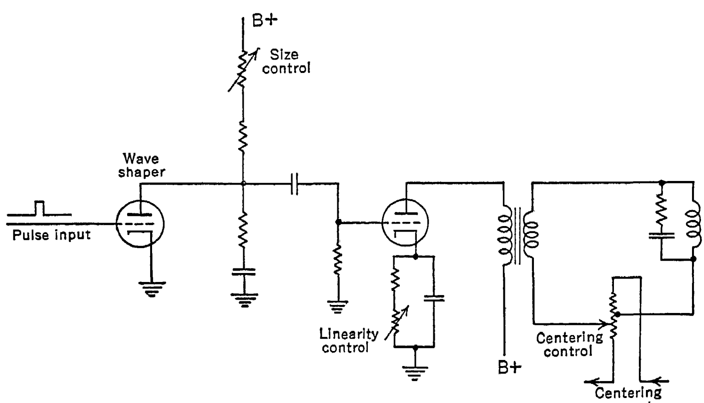


Fig. 14. Field Scanning Circuit

33. POWER SUPPLY

The successful performance of a receiver is largely dependent on its power-supply characteristics. When costs are important, the problems of power supply are among the most difficult which the receiver designer must face.

HEATERS. Coupling. Undesired coupling in the r-f or i-f amplifiers through the heater wiring can be minimized by grounding one side of the heaters. It is usually necessary to provide appropriate by-passing of the heater lead and to interpose occasional r-f chokes.

Heater-cathode Potentials. Except for a few types, most receiving tubes have maximum heater-to-cathode potential ratings in the neighborhood of 100 volts. Since some television circuits employ tubes with cathodes at potentials relative to ground in excess of this rating, it is necessary to provide additional windings on the power transformer when these circuits are used. Where the off-ground cathode is at a-c ground potential, it is customary to connect one side of the heater winding or the center tap to the cathode. If the cathode is not at a-c ground potential, the separate winding may be connected either to the cathode through a high resistor or to a bleeder whose voltage approximates the operating potential of the cathode.

LOW-VOLTAGE B SUPPLY. The two major problems associated with the low-voltage B supply for a television receiver are power-frequency ripple and cross coupling.

Power-frequency Ripple. The entrance of the power-line frequency disturbance in a television picture may cause horizontal bands of light and dark areas, narrowing, widening, and lateral or vertical displacement of parts of the picture, and defocusing. These effects may be caused by the deflection of the cathode-ray-tube beam by the magnetic field of the power transformer, or by insufficient power-frequency filtering in the power supply.

When a receiver is operated from a power line which is non-synchronous with the field scanning, the "hum" pattern drifts upwards or downwards and is very objectionable. It appears essential that commercial receivers be so designed that reception under these conditions remains unimpaired, since programs are even now being relayed over several hundred miles.

As the frequency of the power line at the receiver departs from synchronism with the field scanning, the slowly drifting "hum" pattern jitters and flickers. This is most objectionable in even the slightest amounts. For this reason, "hum" disturbance in receivers intended for operation from power sources whose frequencies are not synchronous with the field scanning must not exceed about 2 per cent.

The technique for securing adequate hum filtering is to provide inductance-capacitance filters in high-current supplies and resistance-capacitance filters in low-current supplies.

Magnetic coupling into the picture tube is most easily cured by placing the power supply at a sufficient distance. When this is impractical, shields of high-permeability material are placed about the tube.

Cross Coupling. Unless precautions are taken, the B supply of a television receiver can be a troublesome source of undesired coupling between the various circuits. Tubes taking high peak currents modulate the B supply, which may disturb circuits that handle low-level signals. To minimize this trouble, multiple power supplies, large filter output capacitors, such as 50 to 100 μf, multisection filters, and separate filters for susceptible circuits are used.

HIGH-VOLTAGE SUPPLIES. The television receiver high-voltage supply provides the power required by the second anode, and, in some cases, the focus electrode, of the cathode-ray tube. The common types rectify power-line frequency voltage, a separately generated sine-wave voltage, or the voltage surge present during the retrace of the line-scanning oscillator. The voltages required range from 2 Kv for small direct-viewed tubes to 30 Kv for projection tubes. The useful current drain from a high-voltage supply may vary between a few microamperes with a dark screen to about 1 ma for a bright picture. Reasonably good voltage regulation is required over the black-to-white current range if noticeable change in picture size and defocusing is to be avoided.

To operate successfully over a number of years, the high-voltage power supply must be designed to withstand the voltage it generates. The spacing around parts at the high potential must be sufficient to prevent sparkover and the formation of corona under unfavorable atmospheric conditions. Insulating materials should be non-carbonizing so that they are not damaged by a single flashover and do not become semi-conducting after periods of service. Insulating paths should be sufficiently long to prevent excessive leakage under conditions of dust and high humidity. Where organic materials must be used, as in a power transformer or paper capacitor, long life can be assured only if these components are impregnated with a good dielectric fluid and are hermetically sealed.

Power-line Frequency Supplies. This type is commonly used where the high voltage required does not exceed 4 or 5 Kv. A single section π filter comprising two capacitors and a resistor is used to remove the power-frequency ripple. The resistor is usually made as large as the voltage regulation requirements will permit and the capacitors as small as will still afford adequate filtering. The capacitors required are usually between about 0.03 and 0.1 μf. Since considerable energy is stored in such capacitors when charged to voltages in excess of a kilovolt, these supplies can be lethal and thus are not suitable for home use unless adequate safety precautions are taken. These include the following:

1. Compartmenting the power supply.
2. Interlocks to prevent access to the power-supply compartment when the power is on.
3. Bleeders capable of discharging the high-voltage capacitors to a safe voltage within a second or less.

4. A substantial connection between the receiver chassis and ground.

Locally Generated Sine-wave Supplies (references 73 and 83). This type is frequently used either where no a-c power line is available or where, for the higher voltages, the cost of the transformer, capacitors, and safety features of a power-line frequency supply are excessive. It comprises a sine-wave oscillator usually operating in the low r-f range, a step-up winding, and one or more vacuum rectifiers in either half-wave or voltage multiplying circuits. The rectifier filaments are customarily powered from the oscillator by windings on the step-up transformer. Adequate filtering is usually obtained with a π filter comprising two capacitors and a resistor, but the capacitors are only a few hundred micromicrofarads. When the energy storage in the capacitors is kept low, this type of supply can be safe even though exposed and usually requires only sufficient shielding to prevent interference by the locally generated wave.

Voltage Surge Supplies (U. S. Patent 2,051,372). This type of power supply rectifies the voltage surge across an inductor when the magnetic field surrounding the inductor is suddenly changed.

The voltage surge type of supply is used in some receivers employing a magnetically deflected picture tube. A high-voltage winding on the line scan output transformer yields the required voltage surge during the scanning retrace. This is rectified and filtered, and the resultant voltage is applied to the second anode of the picture tube. It is necessary to supply added scanning power when the second anode power is thus extracted.

Picture width, when using a high-voltage supply as described above, is adjusted by changing the current through the yoke while not disturbing the currents in the transformer. Otherwise the scanning power and the second anode potential vary together and no significant change in picture width results.

OTHER FORMS OF TELEVISION

By A. V. Loughren

The material of articles 2–33 relates primarily to monochrome, monocular television of sharpness acceptable for home entertainment, for use with power-supply systems of 60-cycle frequency. A change in any of these requirements may affect significantly the design of the entire television system. The more important examples of such changed requirements include: (*a*) television standards of foreign countries; (*b*) theater television; (*c*) color television; (*d*) binocular or stereoscopic television; (*e*) television for special uses (e.g., military, industrial); and (*f*) use of a common transmitter carrier for picture and sound modulations.

34. TELEVISION STANDARDS OF FOREIGN COUNTRIES

With respect to power-supply frequency, the practices of the several countries differ. In addition, since there has been no attempt at international standardization, incidental and unnecessary differences in standards exist. The standards used by the British Broadcasting Corporation (adopted in 1937 and reaffirmed in 1944) illustrate both points. They include:

(*a*) Picture repetition rate: 25 cycles per second (standard power-supply frequency is 50 cycles).
(*b*) Lines per frame: 405.
(*c*) Polarity of picture modulation: positive.
(*d*) Form of picture modulation: amplitude, double side band.
(*e*) Sound modulation: amplitude.
(*f*) Sound carrier location: 3.5 Mc below the picture carrier frequency.

Comparison may be made with the American standards, tabulated in article 10. In most other respects, BBC standards do not differ significantly from those of the United States (references 74 and 75).

35. THEATER TELEVISION

Requirements for theater television differ from those for home television primarily in the following respects:

Highlight Brightness. Motion-picture practice provides highlight brightnesses of 2 to 20 ft-lamberts; the house is dimmed sufficiently to make this acceptable.

Resolution. Pictures projected from 35-mm film have resolution considerably exceeding that of television with a 4-Mc band width. There is some doubt that this factor outweighs the practical advantage of common standards, especially in view of Fig. 1 of article 1.

Operator. Availability of a trained operator.

Picture Size. Motion-picture screens range up to 20 ft in width.

Source of Picture. Provision will probably be required for displaying pictures whether originally picked up for broadcasting or picked up specifically for a chain of theaters. This indicates the desirability of the common standards referred to previously.

36. COLOR TELEVISION

For reasonably faithful reproduction of colored subjects, each geometrical picture element must be represented not by a single intensity, its brightness (as required for monochrome television), but by three separate quantities. These may be the brightnesses of three "primary" colors. Alternatively, one signal may represent the resultant hue, a second the saturation of that hue, and the third, the brightness. Color measurement and specification may be done by either of these alternatives.

Color reproduction systems are divided into additive and subtractive systems. In an additive system, for each picture element an individual stream of light energy in each of the primary colors reaches the eye of the observer. If the light as originally generated is white, a major portion of it is discarded in the filters, which transform it to light of the primary colors. In a subtractive system, the originally produced white light is modified individually for each picture element by subtracting only the unwanted color components. The subtractive system consequently shows an efficiency in the use of a white-light source which is several times that of an additive system. In color photography, subtractive processes such as Technicolor and Kodachrome have found much greater acceptance than the additive processes such as Autochrome and Finlay. In color television, on the other hand, only systems of the additive type have been developed sufficiently to promise practical utility.

For additive systems, desirable reproduction primaries are red, green, and blue, individually chosen as compromises between purity (or saturation to increase the range of producible colors) and transmission loss from white light. Colors equivalent to the transmissions through Wratten filters numbers 47, 58, and 25, when the light source is the International Committee on Illumination's Illuminant "C," have found some acceptance as the red, green, and blue primaries.

METHODS OF TRANSMISSION. Numerous methods of transmission of color television signals have been proposed. Of the several ways of classifying these methods a classification by the time characteristics of the signal seems most important. On this basis, the systems may be classified as:

1. Simultaneous systems, in which the three elements of information required to describe a single picture element are transmitted simultaneously.

2. Sequential systems, in which the three elements of information are transmitted successively. The sequential systems which have been proposed fall into the following subclasses:

(a) *Field Sequential.* A complete picture field is transmitted in one color, followed by successive fields in the remaining colors. In a three-color system with 2 : 1 interlace, six fields must elapse before a picture which is complete both geometrically and in color may be obtained.

(b) *Line Sequential.* A line of one color is transmitted, followed by successive lines in the remaining colors, and the cycle repeats. In this system if the number of colors is an integral submultiple of the number of lines in a complete picture, a given picture line (for example, the eleventh) will always be repeated in the same color unless the color sequence switching is momentarily altered at the end of a frame to produce a new phasing for the next frame. With three colors it is difficult to avoid in a line sequential system a "crawling" tendency in the produced image which has usually been exhibited in systems with orders of interlace greater than 2 : 1.

(c) *Dot Sequential.* The three bits of information describing an individual picture element are transmitted in immediate succession, after which transmission of information for the next picture element takes place. In this system a dot pattern somewhat similar to that of a halftone engraving appears superposed on the colored portions of the picture. It is interesting to note that dot sequential systems are closer in their characteristics to simultaneous systems than to the other sequential systems.

Color television systems may be also classified in accordance with the quantities explicitly represented by the elements of the transmitted signals. Among the many possibilities are the following:

1. Intensities of individual primary colors.
2. Composite intensity (or, alternatively, visual brightness) plus two auxiliary signals representing, for example, the difference between the apparent brightness and the red and blue intensities respectively.
3. Composite intensity, hue, and saturation.

PHYSIOLOGICAL REQUIREMENTS. Resolution. It has been shown that, in a color picture, the apparent resolution is not appreciably impaired when the blue image is severely defocused; it is also known that moderate defocusing of the red image is permissible. Further, the use of a common signal to represent the fine detail for all three colors can be demonstrated as causing negligible impairment of picture quality, when practiced in reasonable amounts.

Since brightness is largely determined by the green content of a color, it appears from the foregoing facts that resolution may be effectively preserved, and yet the frequency band effectively conserved, by a system of transmission in which brightness is transmitted with a band of several megacycles, while hue and saturation are transmitted with relatively narrow bands. If this practice is applied to a system in which separate components of the radiated signal represent the intensities of the primary colors, the low-frequency portion of each color component of the signal will be derived individually (to represent the distribution of light of its corresponding color in the subject) while the high-frequency portion will be identical in all three colors. This common high-frequency signal has been called "mixed highs."

No studies comparable to that of reference 4 have been made for color pictures; in the absence of such information, it seems probable that resolution comparable to that of monochrome television is desirable, *especially for the brightness component.*

Flicker. The apparent brightness of a color image is determined largely by the green component. If the entire image (or an interlaced image field) is produced in each color, the flicker performance is essentially that given by considering the repetition rate of the green component only.

Color Range. The available color reproduction range of a television system using the reproducing primaries suggested above is comparable to the best ranges obtained by commercial color reproduction processes of other sorts such as color film and multicolor printing. Such a range appears adequate.

COMPATIBILITY. When a color television system is put into use in an area already provided with a monochrome television service, the question of compatability arises. A color television system is compatible with a particular monochrome television system when color television signals radiated by the color system may be received as monochrome images of acceptable quality on the receivers of the monochrome system without modification to such receivers. The possession of this characteristic by a color television system contributes importantly to making the commercial introduction of the color system easy since:

1. Color transmissions may be started by individual stations of the monochrome television service as soon as the stations are equipped for color pickup, with no loss of audience or impairment of audience satisfaction.
2. Those viewers who are immediately interested in receiving color may purchase new color receivers with assurance that, since the audience for color broadcasts includes both themselves and the monochrome audience, program service will develop rapidly.

For a color system to be compatible with a monochrome system it must employ essentially the same transmission standards as the monochrome system. Any changes in the standards either must be small enough in amount to be without effect on operation of the monochrome receivers or must be in the nature of additions to the signal of a sort which will go undetected by the monochrome receivers. Dot sequential systems, because of their superior potentialities for band-width economy, seem more likely to be capable of operation compatible with the present FCC monochrome standards with a satisfactory grade of reproduced picture than the systems using slower color sequence rates.

TRANSMITTER. Color television transmitters differ necessarily from monochrome transmitters in requiring the use of a color camera and possibly additional control signals for color synchronizing information, etc. In other respects, however, they may be essentially similar to a monochrome transmitter. The detailed requirements for the camera and the additional control signals are dependent on the particular color television system considered.

RECEIVER. A receiver for color television differs necessarily from a receiver for monochrome television only in the substitution of a display capable of giving color reproduction and the addition of circuits to convert the output of the receiver's detector into signals appropriate to the particular display. The circuits needed are those used to synchronize and phase the effective color of the display to correspond to the color of the bit of picture information being supplied to the display at the corresponding instant. If the system is field sequential so that the color switching rate is once per field, the circuits may control the synchronizing and phasing of a motor driving a color wheel—a disk provided with a series of sectors in the successive colors. In an all-electronic display the color signals must perform a corresponding function; in the electronic arrangements the rate of switching from one color to the next may be arranged to accommodate either dot sequential, line sequential, or field sequential systems.

See also references 1, 5, 25, 76, 77, 85–92.

37. BINOCULAR TELEVISION

In direct viewing of a scene, the images formed in an observer's right and left eyes differ. Depth perception is based, in part, on this difference, as discussed in Section 14. Presentation of suitably different reproduced images to an observer's two eyes will in many cases enhance the illusion of solidity in a reproduced picture.

Binocular or stereoscopic presentation to a viewer must provide that each eye sees only the picture intended for it. Among the means used for this purpose have been (1) barriers and lenses to direct each eye correctly; (2) alternate display of the right and left images with spectacles worn by the viewer containing a synchronized mechanical shutter; (3) alternate display of a red right image and a green left image, with viewer spectacles having no green transmission to the right eye and no red transmission to the left eye; and (4) alternate display of a right image with horizontally polarized light and a left image with vertically polarized light, with complementary viewer spectacles.

For the three-dimensional illusion to be most effective, the spacing between the positions of the camera in taking the images should be equal to the average ocular separation (unless the image is to be magnified in reproduction). The presentation of separate pictures for the two eyes requires transmission of twice the information (and, hence, twice the band width) required for monocular presentation.

The need for special viewing devices has prevented any wide interest in binocular television.

38. TELEVISION FOR SPECIAL SERVICES

Proposals for industrial and for military uses of television have often presented requirements radically different from those of television broadcasting for home entertainment. Examples of these unusual requirements have included: (1) effective protection against jamming; (2) secrecy; (3) severe size and weight limitations; (4) unattended transmitter and camera operation; (5) the high importance accorded to reliability; (6) the usual relatively unfavorable operating conditions (as compared to household use) encountered in many forms of industrial and military apparatus design. Requirements of this sort may affect basically the design of the television system, in addition to their obvious effect on the detailed design of the apparatus. For examples of designs for military purposes, see reference 78.

39. DIPLEXING OF PICTURE AND SOUND

Use of a single carrier for both picture and sound modulations is of interest because of possible simplification of receivers and decrease of radio spectrum space resulting from it. Possible approaches include: (*a*) use of different (and mutually non-interfering) forms of modulation for the two signals—for example, amplitude modulation for the picture, with frequency modulation for the sound; (*b*) sharing of time between picture and sound signals.

Method (*a*) has not been tried widely, probably in view of the general use of vestigial sideband picture transmission, with its inherent introduction of picture frequency modulation sidebands representing all but the lowest frequencies and the consequent likelihood of cross-talk between the two signals in a receiver.

Method (*b*) has been proposed on several occasions and has had some laboratory and field trials. It is found that the frequency of the intervals in which sound signals are transmitted is between two and three times the highest sound modulation frequency which can be successfully transmitted, and that the resulting sound-signal-to-noise ratio is dependent on the fraction of the total time used for sound and on the details of the sound modulation process. The sound intervals must be so located relative to the picture signals, in time, as to produce no visible effect; they may, therefore, be placed between the synchronizing signal and the start of the picture information for each scanning line. See reference 79 for a more complete discussion and bibliography.

REFERENCES

1. "Optics," Section 14 of this handbook.
2. Wilson, J. C., *Television Engineering*, Pitman (1937).
3. Engstrom, E. W., A Study of Television Image Characteristics, *Proc. I.R.E.*, Vol. 21, No. 12, pp. 1631–1651 (December 1933), and Vol. 23, No. 4, pp. 295–310 (April 1935).
4. Baldwin, M. W., The Subjective Sharpness of Television Images, *Proc. I.R.E.*, Vol. 28, No. 10, pp. 458–468 (October 1940).
5. Hardy, A. C., *Handbook of Colorimetry*, Technology Press (1936).
6. Mertz, P., and F. Gray, A Theory of Scanning, *Bell Sys. Tech. J.*, Vol. 13, No. 3, pp. 464–515 (July 1934).
7. Wheeler, H. A., and A. V. Loughren, The Fine Structure of Television Images, *Proc. I.R.E.*, Vol. 26, No. 5, pp. 540–575 (May 1938).
8. Campbell, G. A., and R. M. Foster, Fourier Integrals for Practical Applications, *Bell Telephone System Monograph* B584.
9. Wheeler, H. A., The Interpretation of Amplitude and Phase Distortion in Terms of Paired Echoes, *Proc. I.R.E.*, Vol. 27, No. 6, pp. 359–385 (June 1939).
10. Robinson, D. M., The Supersonic Light Control, *Proc. I.R.E.*, Vol. 27, No. 8, pp. 483–486 (August 1939).
11. Sieger, J., Television Receivers Using the Scophony Optical Scanning System, *Proc. I.R.E.*, Vol. 27, No. 8, pp. 487–492 (August 1939).
12. Wikkenhauser, G., Synchronization of Scophony Television Receivers, *Proc. I.R.E.*, Vol. 27, No. 8, pp. 492–496 (August 1939).
13. Lee, H. W., Optical Design of Television Receiver Using Moving Scanners, *Proc. I.R.E.*, Vol. 27, No. 8, pp. 496–500 (August 1939).
14. Rosenthal, A. H., A System of Large-screen Television Reception Based on Certain Electron Phenomena in Crystals, *Proc. I.R.E.*, Vol. 28, No. 5, pp. 203–212 (May 1940).
15. Donal, J. S., Jr., Cathode-ray Control of Television Light Valves, *Proc. I.R.E.*, Vol. 31, No. 5, pp. 195–208 (May 1943).
16. Donal, J. S., Jr., and D. B. Langmuir, A Type of Light Valve for Television Reproduction, *Proc. I.R.E.*, Vol. 31, No. 5, pp. 208–214 (May 1943).
17. Wendt, K. R., and G. L. Fredendall, Automatic Frequency and Phase Control of Synchronization in Television Receivers, *Proc. I.R.E.*, Vol. 31, No. 1, pp. 7–15 (January 1943).
18. Poch, W. J., and D. W. Epstein, Partial Suppression of One Side Band in Television Reception, *Proc. I.R.E.*, Vol. 25, No. 1, pp. 15–31 (January 1937).
19. Hollywood, J. M., Single-Sideband Filter Theory with Television Applications, *Proc. I.R.E.*, Vol. 27, No. 7, pp. 457–472 (July 1939).
20. Nergaard, L. S., A Theoretical Analysis of Single-Sideband Operation of Television Transmitters, *Proc. I.R.E.*, Vol. 27, No. 10, pp. 666–677 (October 1939).
21. Goldman, Stanford, Television Detail and Selective-Sideband Transmission, *Proc. I.R.E.*, Vol. 27, No. 11, pp. 725–732 (November 1939).
22. Kallman, H. E., and R. E. Spencer, Transient Response of Single-Sideband Systems, *Proc. I.R.E.*, Vol. 28, No. 12, pp. 557–561 (December 1940).
23. Singer, C. P., A Mathematical Appendix to Transient Response of Single-Sideband Systems, *Proc. I.R.E.*, Vol. 28, No. 12, pp. 561–563 (December 1940).
24. Wheeler, H. A., The Solution of Unsymmetrical-Sideband Problems with the Aid of the Zero-Frequency Carrier, *Proc. I.R.E.*, Vol. 29, No. 8, pp. 446–458 (August 1941).
25. Goldmark, Dyer, Piore, and Hollywood, Color Television Part I, *Proc. I.R.E.*, Vol. 30, No. 4, pp. 162–182 (April 1942); Goldmark, Piore, Hollywood, Chambers, and Rives, Color Television Part II, *Proc. I.R.E.*, Vol. 31, No. 9, pp. 465–478 (September 1943).
26. Zworykin, V. K., The Iconoscope, *Proc. I.R.E.*, Vol. 22, No. 1, pp. 16–32 (January 1934).
27. Bedford, A. V., and J. P. Smith, A Precision Television Synchronizing Signal Generator, *RCA Rev.*, Vol. V, No. 1, pp. 51–68 (July 1940).
28. Fink, D. G. (editor), *Television Standards and Practice—NTSC*, McGraw-Hill.
29. Mertz, P., Television—The Scanning Process, *Proc. I.R.E.*, Vol. 29, No. 10, pp. 529–537 (October 1941).
30. Rose, Weimer, and Law, The Image Orthicon—A Sensitive Television Pickup Tube. *Proc. I.R.E.*, Vol. 34, No. 7, pp. 424–432 (July 1946).
31. Beers, G. L., The Focusing Viewfinder Problem in Television Cameras, *Proc. I.R.E.*, Vol. 31, No. 3, pp. 100–106 (March 1943).
32. Nicoll, F. H., A New Chemical Method of Reducing the Reflection of Glass, *RCA Rev.*, Vol. 6, No. 3, pp. 287–301 (January 1942).
33. De Vore and Iams, Some Factors Affecting the Choice of Lenses for Television Cameras, *Proc. I.R.E.*, Vol. 28, No. 8, pp. 369–374 (August 1940).
34. Larsen and Gardner, Image Dissector, *Electronics*, Vol. 12, p. 24 (October 1939).
35. Monfort and Somers, Measurement of the Slope and Duration of Television Synchronizing Impulses, *RCA Rev.*, Vol. 6, No. 3, pp. 370–389 (January 1942).
36. Maloff, I. G., Gamma and Range in Television, *RCA Rev.*, Vol. 3, No. 4, pp. 409–417 (April 1939).
37. Bamford, H. S., A New Television Film Projector, *Electronics*, Vol. 11, p. 25 (July 1938.)
38. Engstrom, Beers, and Bedford, Application of Motion Picture Film to Television, *RCA Rev.*, Vol. 4, No. 1, pp. 48–81 (July 1939).
39. Bedford and Fredendall, Transient Response of Multistage Video Frequency Amplifiers, *Proc. I.R.E.*, Vol. 27, No. 4, pp. 277–284 (April 1939).
40. Bedford and Fredendall, Analysis, Synthesis, and Evaluation of the Transient Response of Television Apparatus, *Proc. I.R.E.*, Vol. 30, No. 10, pp. 440–457 (October 1942).
41. Loughlin, B. D., A Phase Curve Tracer for Television, *Proc. I.R.E.*, Vol. 29, No. 3, pp. 107–115 (March 1941).
42. Kell, Bedford, and Kozanowski, A Portable High Frequency Square Wave Oscillograph for Television, *Proc. I.R.E.*, Vol. 30, No. 10, pp. 458–464 (October 1942).
43. Townsend, C. L., Contemporary Problems in Television Sound, *Proc. I.R.E.*, Vol. 31, No. 1, pp. 3–7 (January 1943).

44. Shelby and See, Field Television, *RCA Rev.*, Vol. 7, No. 1, pp. 77–93 (March 1946).
45. Strieby and Wentz, Television Transmission over Wire Lines, *Bell Sys. Tech. J.*, Vol. 20, No. 1, pp. 62–81 (January 1941).
46. Strieby and Weis, Television Transmission, *Proc. I.R.E.*, Vol. 29, No. 7, pp. 371–381 (July 1941).
47. Hansell, C. W., Radio Relay Systems, *Proc. I.R.E.*, Vol. 33, No. 3, pp. 156–168 (March 1945).
48. Friis, H. T., A Note on a Simple Transmission Formula, *Proc. I.R.E.*, Vol. 34, No. 5, pp. 354–356 (May 1946).
49. Parker, W. N., A Unique Method of Modulation for High Fidelity Television Transmission, *Proc. I.R.E.*, Vol. 26, No. 8, pp. 946–962 (August 1938).
50. PJHA Nordlohne, Experimental Short Wave Broadcasting Station PCJ, *Phillips Tech. Rev.*, Vol. 3, No. 1, pp. 17–27 (January 1938).
51. Kell and Fredendall, Selective Side Band Transmission in Television, *RCA Rev.*, Vol. 4, No. 4, pp. 425–440 (April 1940).
52. Brown, G. H., A Vestigial Side Band Filter, *RCA Rev.*, Vol. 5, No. 3, pp. 301–326 (January 1941).
53. Buzalski, T. J., A Method of Measuring the Degree of Modulation of a Television Signal, *RCA Rev.*, Vol. 7, No. 2, pp. 265–271 (June 1946).
54. Lindenblad, N. E., Television Transmitting Antenna, *RCA Rev.*, Vol. 3, No. 4, pp. 387–408 (April 1939).
55. Carter, P. S., Simple Television Antennas, *RCA Rev.*, Vol. 4, No. 2, pp. 168–185 (October 1939).
56. Janes, R. B., and W. A. Hickok, Recent Developments in the Design Characteristics of the Iconoscope, *Proc. I.R.E.*, Vol. 27, No. 9, p. 535 (September 1939).
57. Johnson, E. O., Development of an Ultra Low Loss Transmission Line for Television, *RCA Rev.*, Vol. VII, No. 2, pp. 272–280 (June 1946).
58. Herold, E. W., Local Oscillator Radiation and Its Effect on Television Picture Contrast, *RCA Rev.*, Vol. VII, No. 1, pp. 32–53 (March 1946).
59. Herold, E. W., Superheterodyne Converter System Considerations in Television Receivers, *RCA Rev.*, Vol. IV, No. 3, pp. 324–337 (January 1940).
60. Herold, E. W., The Operation of Frequency Converters and Mixers for Superheterodyne Reception, *Proc. I.R.E.*, Vol. 30, No. 2, pp. 84–103 (February 1942).
61. Jones, M. C., Grounded-Grid Radio-Frequency Voltage Amplifiers, *Proc. I.R.E.*, Vol. 32, No. 7, pp. 423–429 (July 1944).
62. Seeley, S. W., Effect of the Receiving Antenna on Television Reception Fidelity, *RCA Rev.*, Vol. II, No. 4, pp. 433–441 (April 1938).
63. Herold, E. W., An Analysis of the Signal to Noise Ratio of Ultra High Frequency Receivers, *RCA Rev.*, Vol. VI, No. 3, pp. 302–331 (January 1942).
64. Mountjoy, G., Television Signal-Frequency Circuit Considerations, *RCA Rev.*, Vol. IV, No. 2, pp. 204–230 (October 1939).
65. Tyson, B. F., A Preselector Circuit for Television, *Electronics*, Vol. 13, pp. 23–25 (November 1940).
66. Nyquist, H., Certain Topics in Telegraph Transmission Theory, *Trans. A.I.E.E.*, Vol. 47, pp. 617–644 (April 1928).
67. Mountjoy, G., Simplified Television IF Systems, *RCA Rev.*, Vol. IV, No. 3, pp. 299–309 (January 1940).
68. Freeman, R. L., The Use of Feedback to Compensate for Input-Capacitance Variations with Grid Bias, *Proc. I.R.E.*, Vol. 26, No. 11, pp. 1360–1366 (November 1938).
69. Wheeler, H. A., Wide-Band Amplifiers for Television, *Proc. I.R.E.*, Vol. 27, No. 7, pp. 429–438 (July 1939).
70. Zworykin, V. K., and G. A. Morton, *Television* (a book), Chapter 15, p. 465; Chapter 17, p. 524, Wiley (1940).
71. Puckle, O. S., *Time Bases*, Chapter V, pp. 64–66, 68–69; Chapter VIII, pp. 115–122, Wiley (1943)
72. Maloff, I. G., and D. W. Epstein, *Electron Optics in Television*, Chapter 13, p. 255, McGraw-Hill.
73. Schade, O. H., Radio-Frequency Operated High-Voltage Supplies for Cathode-Ray Tubes, *Proc. I.R.E.*, Vol. 31, No. 4, pp. 158–163 (April 1943).
74. Blumlein, A. D., The Marconi-E.M.I. Television System: Part I. The Transmitted Wave-Form, *J. I.E.E.*, Vol. 83, No. 504, pp. 758–766 (December 1938).
75. *Report of the Television Committee, 1943*, Right Hon. Lord Hankey, Chairman, His Majesty's Stationery Office, London.
76. Judd, Deane B., Introduction to Color, from *Symposium on Color*, American Society for Testing Materials, pp. 1–12.
77. Friedman, Joseph S., *History of Color Photography*, American Photographic Publishing Company.
78. Marshall and Katz, Television Equipment for Guided Missiles, *Proc. I.R.E.*, Vol. 34, No. 6, pp. 375–401 (June 1946).
79. Fredendall, Schlesinger, and Schroeder, Transmission of Television Sound on the Picture Carrier, *Proc. I.R.E.*, Vol. 34, No. 2, pp. 49–61 (February 1946).
80. Sziklai, G. C., Current Oscillator for Television, *Electronics*, Vol. 19, No. 9, pp. 120–123 (September 1946).
81. Zworykin and Morton, *Television*, Chapter 15, pp. 471–473, Wiley (1940).
82. Friend, A. W., Television Deflection Circuits, *RCA Rev.*, Vol. VIII, No. 1 (March 1947).
83. Mautner, R. S., and O. H. Schade, Television High Voltage R-F Supplies, *RCA Rev.*, Vol. VIII, No. 1 (March 1947).
84. Wendt, K. R., The Television D-C Component, *RCA Rev.*, Vol. 9, No. 1, pp. 85–111 (March 1948).
85. Evans, Ralph M., *An Introduction to Color*, Wiley (1948).
86. An Experimental Simultaneous Color-Television System: Kell, R. D., Part I, Introduction; Sziklai, G. C., Ballard, R. C., and Schroeder, A. C., Part II, Pickup Equipment; Wendt, K. R., Fredendall, G. L., and Schroeder, A. C., Part III, Radio Frequency and Reproducing Equipment; *Proc. I.R.E.*, Vol. 35, No. 9, pp. 861–875 (September 1947).
87. Battison, John H., Color Television Transmission Systems, *Tele-Tech*, Vol. 8, No. 10, pp. 18–20 (October 1949).
88. Report of FCC Color TV Demonstrations at Washington, *Tele-Tech*, Vol. 8, No. 11, pp. 24–26 (November 1949).
89. Loomis, Franklin, Highlights of FCC Color-TV Demonstrations, *Tele-Tech*, Vol. 8, No. 12, pp. 18–19 (December 1949).
90. New Directions in Color Television, *Electronics*, Vol. 22, No. 12, pp. 66–72 (December 1949).
91. Boothroyd, Wilson, Dot Systems of Color Television, Part I, *Electronics*, Vol. 22, No. 12, pp. 88–92 (December 1949); Part II, *Electronics*, Vol. 23, No. 1, pp. 96–99 (January 1950).
92. RCA Laboratories, A Six-Megacycle Compatible High-Definition Color Television System, *RCA Rev.*, Vol. 10, No. 4, pp. 504–524 (December 1949).

SECTION 21

ELECTRONIC CONTROL EQUIPMENT

BY

B. J. DALTON

ELECTRONIC CONTROL EQUIPMENT

By B. J. Dalton

FUNDAMENTAL ELECTRONIC POWER CIRCUITS

The fundamental purpose of any electronic rectifier is to convert alternating current into direct current. Therefore, it can be considered as a d-c power supply, the same as a battery or a motor-generator set. There are, however, two important differences between rectifiers and generators or batteries:

1. Batteries supply a smooth d-c voltage output; generators have a number of commutator segments so that the instantaneous voltage at the brushes is nearly constant; rectifiers, on the other hand, generally consist of a relatively small number of phases and rectifying elements and therefore almost all of them have a considerable amount of ripple voltage in the output. The ripple voltage must be considered in many applications, and in particular those involving less than 6-phase rectification, because the ripple voltage may produce a current ripple in some loads which will cause excessive heating not only in the load but also in the rectifier transformer and the rectifying elements. A highly inductive load is a very desirable rectifier load, because the inductance smooths out most of the current ripple. Resistance loads and particularly counter emf or capacitive loads will result in high rms currents which cause additional heating. In some counter-emf type loads it may be desirable to include sufficient reactance in the rectifier output circuit to limit the peak current to a reasonable value. This obviates the necessity for excessive rectifying element and transformer sizes and also minimizes the additional heating which would otherwise be present in the load.

2. Batteries are inherently energy-storage devices and have a constant output voltage; generators not only have a small amount of energy stored mechanically in their rotor but also are usually driven by an a-c motor the speed of which is reasonably independent of a-c line voltage; thus the generator has a constant output voltage. Rectifiers, on the other hand, have no inherent energy storage, and the output voltage at any instant is directly proportional to the a-c input voltage. Therefore, if a constant d-c output voltage is desirable, either the a-c input voltage must be regulated or a regulating means must be provided for the output voltage.

The efficiency of electronic rectifiers is determined by the losses in the rectifying elements themselves together with transformer losses, cathode heating losses, and miscellaneous auxiliary losses. The power factor of a rectifier is determined by the type of load and number of phases and for a controlled rectifier by the amount of phase retard as well. An inductive load will give the highest power factor; a counter-emf load will result in the lowest power factor. For a given type of load, the power factor is inversely proportional to the amount of phase retard, the rectifier being essentially a constant-kva load on the power line.

One or more of the following factors will govern the choice of the type of rectifier for a specific application:

1. The magnitude of the required d-c power.
2. The magnitude of the required d-c voltage.
3. The magnitude of the required d-c current.
4. The required degree of freedom from a-c voltage ripple in the d-c output voltage or current.
5. The effect of the connected load on the rectifying elements.
6. The number of phases and the type of connection available from the a-c power supply.
7. The magnitude of the a-c power supply voltage.
8. The necessity for adjusting the output voltage.
9. The necessity for regulating the output voltage at a specific value.
10. Physical size.
11. Cost.

Electronic rectifiers may be classified in accordance with their controllability as (1) noncontrolled, (2) grid-controlled, and (3) ignitron or pool-type-controlled rectifiers. (These types will be discussed in greater detail later.) Also electronic rectifiers may be classified in accordance with the type of circuit connection used.

The proposed AIEE standards for pool-cathode mercury-arc power converters include a list of 36 standard rectifier circuit connections. Table I, however, illustrates eight circuit connections which are representative of a large number of applications, and in particular those applications in the low- and medium-power field. These circuits and their uses are described below.

1. RECTIFIER CIRCUITS AND APPLICATIONS

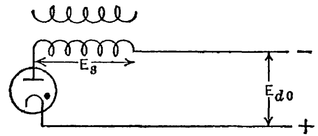
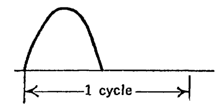
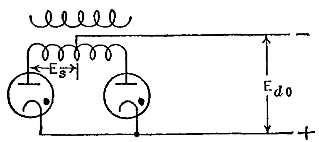
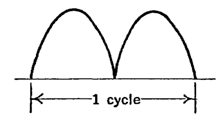
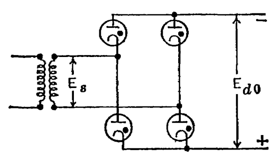
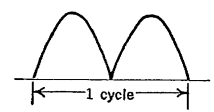
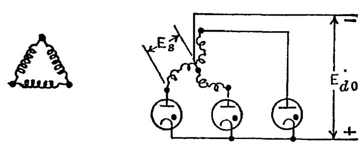
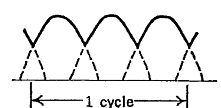
Half-wave rectifiers are generally limited in their application to low power circuits. Typical half-wave rectifier applications are: (1) D-c power supplies for small electronic amplifiers. In this application a filter circuit is used to store energy during the half cycle in which the rectifier element is conducting so that a reasonably smooth output voltage can be obtained. (2) D-c power for operating d-c coils of magnetic relays. (3) Charging circuits for capacitors. (4) D-c power supplies for the armature power for small d-c motors. (5) Battery charging equipment. Half-wave rectifiers are generally used because of their simplicity and relatively low cost. Furthermore the half-wave rectifier circuit can be operated directly from a single-phase power supply, without an anode transformer, if the load is designed to match the obtainable output voltage. The half-wave circuit, however, is undesirable in some respects because the energy delivered to the load in an entire cycle must be obtained during one-half cycle. This reduces the transformer utilization (if a transformer is used) and also results in high peak currents in the rectifying element. In addition, this type of rectifier is generally unsuited for application to highly inductive (i.e., iron-core inductance) loads. The time constant of an inductive load is usually sufficiently long so that current will not build up during the half cycle in which the tube can conduct. The small amount of energy that is transferred into the load is inverted during the non-conducting half cycle. Usually the average current is about 10 to 20 per cent of that which would be expected in a pure resistance. A capacitor or a rectifier tube is sometimes connected across an inductive load to prevent the energy stored in the inductance from being transferred to the a-c supply during the normally non-conducting half cycle. Thus the current can build up over a period of time and finally reach a steady-state value.

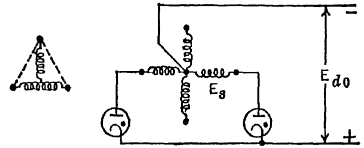
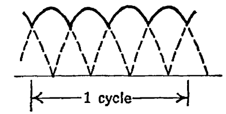
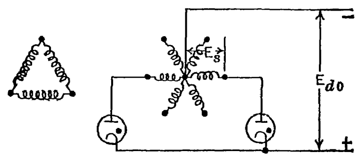
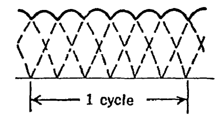
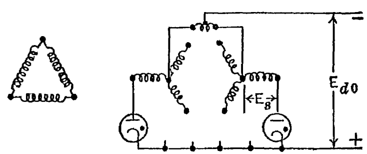
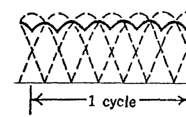
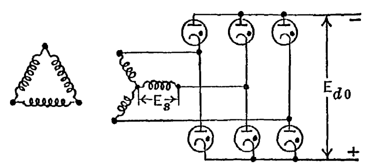
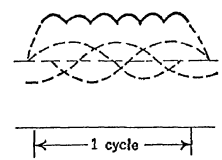
The diametric (full-wave) rectifier is used in a large number of low-power applications. Typical applications are: (1) d-c power supplies for other electronic equipment; (2) d-c power supplies for magnetic clutches, magnetic chucks, and lifting magnets; (3) battery chargers; (4) d-c power for supplying the fields of d-c motors and generators; (5) d-c power for the armature circuit of d-c motors. This rectifier is a relatively simple and inexpensive unit. In some applications, however, its usefulness is limited by the amount of ripple present in the output voltage. The desirability of obtaining larger amounts of power from polyphase power supplies usually limits its use to applications involving less than 5 kw of d-c power.

The primary advantage of the diametric, double-way (bridge) rectifier is its ability to supply high voltages. The peak inverse voltage across the rectifying element is only half of what it would be in a diametric (full-wave) circuit. Thus a diametric, double-way rectifier can be designed to deliver twice the d-c voltage that a diametric rectifier can deliver, assuming the same rectifying elements in both cases. It also results in high transformer utilization. This rectifier circuit has another advantage in that it can be operated directly from a single-phase power supply without an anode transformer, provided that the output voltage obtained this way is of a suitable value.

Polyphase rectifiers are generally used whenever large amounts of power are required. The selection of a particular polyphase rectifier is largely a matter of obtaining the desired output voltage and current from existing or standard rectifying elements. For example, if rectifying elements of 5-amp capacity each were available and a 15-amp d-c output were required, a delta 3-phase wye rectifier circuit could be chosen. Likewise if 20 amp of direct current were required using the same rectifying elements, a Scott 4-phase cross rectifier circuit could be selected. A delta 6-phase star rectifier circuit could be selected to obtain 30 amp of direct current from the same rectifying elements. However, it might be that the use of such a circuit as the delta 6-phase star would place a severe duty on a particular rectifying element either from the standpoint of the rms current or from the standpoint of the peak current. For example, if an excessive peak or rms rectifier current would exist in a delta 6-phase star circuit, the delta 6-phase double-wye circuit could be chosen. This would reduce the rms and peak currents by a ratio of almost 2 to 1. It is sometimes possible by careful coordination of power supply connections and load voltage ratings to use a delta 6-phase wye double-way rectifier to eliminate the rectifier transformer. Double-way rectifiers, as mentioned before, will deliver higher d-c voltages, for a given peak inverse voltage across the rectifying elements, than single-way rectifiers.

Table 1

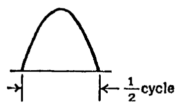
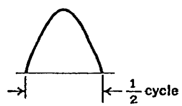
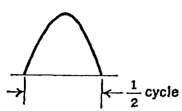
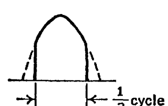
Circuit Nomenclature *	Connection Diagram	Output Voltage Wave Form	Secondary Voltage E_S (RMS) †	Output Ripple: Frequency (Principal Component)	Output Ripple: RMS, volts	Peak Inverse Voltage
Half-wave			$2.22 \times E_{d0}$	Line	$1.11E_{d0}$	$3.14E_{d0}$
Diametric (full-wave)			$1.11 \times E_{d0}$	2 × line	$0.471E_{d0}$	$3.14E_{d0}$
Diametric double-way (bridge)			$1.11 \times E_{d0}$	2 × line	$0.47E_{d0}$	$1.57E_{d0}$
Delta, three-phase, wye			$0.855 \times E_{d0}$	3 × line	$0.18E_{d0}$	$2.09E_{d0}$

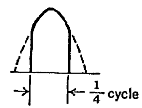
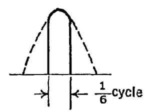
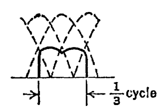
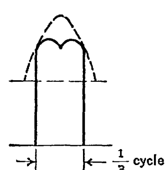
Scott, four-phase cross			$0.787 \times E_{d0}$	4 × line	$0.094E_{d0}$	$2.22E_{d0}$
Delta, six-phase, star			$0.741 \times E_{d0}$	6 × line	$0.04E_{d0}$	$2.09E_{d0}$
Delta, six-phase, double wye			$0.855 \times E_{d0}$	6 × line	$0.04E_{d0}$	$2.09E_{d0}$
Delta, six-phase, wye, double-way			$0.427 \times E_{d0}$	6 × line	$0.04E_{d0}$	$1.047E_{d0}$

* Proposed AIEE standard.

† Neglects voltage loss due to transformer reactance, tube drop, and lag in firing angle.

Table 1—*Continued*

Circuit Nomenclature *	Tube Current Wave Form (Resistance Load)	Rectifying Element Current (Resistance Load)			Transformer Rating			
					Resistance Load		Inductive Load	
		Avg.	Rms	Peak	Total Secondary Kva	Total Primary Kva	Total Secondary Kva	Total Primary Kva
Half-wave		1.0 × I_{d0}	1.57 × I_{d0}	3.14 × I_{d0}	3.49 × d-c kw	2.69 × d-c kw †	Not generally suitable for inductive load	
Diametric (full-wave)		0.5 × I_{d0}	0.786 × I_{d0}	1.57 × I_{d0}	1.75 × d-c kw	1.235 × d-c kw	1.57 × d-c kw	1.11 × d-c kw
Diametric double-way (bridge)		0.5 × I_{d0}	0.786 × I_{d0}	1.57 × I_{d0}	1.235 × d-c kw	1.235 × d-c kw	1.11 × d-c kw	1.11 × d-c kw
Delta, three-phase, wye		0.333 × I_{d0}	0.587 × I_{d0}	1.21 × I_{d0}	1.508 × d-c kw	1.23 × d-c kw	1.481 × d-c kw	1.209 × d-c kw

Scott, four-phase, cross		$0.250 \times I_{d0}$	$0.502 \times I_{d0}$	$1.11 \times I_{d0}$	$1.57 \times$ d-c kw	$1.195 \times$ d-c kw	$1.57 \times$ d-c kw	$1.195 \times$ d-c kw
Delta, six-phase, star		$0.167 \times I_{d0}$	$0.408 \times I_{d0}$	$1.05 \times I_{d0}$	$1.814 \times$ d-c kw	$1.283 \times$ d-c kw	$1.814 \times$ d-c kw	$1.283 \times$ d-c kw
Delta, six-phase, double wye		$0.167 \times I_{d0}$	$0.293 \times I_{d0}$	$0.605 \times I_{d0}$	$1.481 \times$ d-c kw	$1.047 \times$ d-c kw	$1.481 \times$ d-c kw	$1.047 \times$ d-c kw
Delta, six-phase, wye, double-way		$0.333 \times I_{d0}$	$0.579 \times I_{d0}$	$1.05 \times I_{d0}$	$1.074 \times$ d-c kw	$1.047 \times$ d-c kw	$1.047 \times$ d-c kw	$1.047 \times$ d-c kw

* Proposed AIEE standard.

† This is a theoretical value; actual kva will be higher because of transformer exciting current.

Sometimes the requirement of low ripple currents governs the choice of the rectifier circuit, rather than the current capacity. For very large rectifiers above 1000 kw (total installed capacity), 12 or more phase rectifiers may be used to minimize the effect of rectifier harmonics on telephone lines.

EXPLANATION OF TABLE 1. Table 1 consists of four distinct sections appearing from left to right as follows: (1) circuit nomenclature and connection diagrams; (2) output voltage wave shapes, required transformer-secondary voltage ratings, and output-voltage ripple values; (3) rectifying element current wave shapes (for resistance load), rectifying-element current values (for resistance load), and rectifying-element peak inverse voltages; (4) rectifier-transformer kva ratings, for both resistive and inductive loads. All data are tabulated in terms of the theoretical no load d-c output voltage (E_{d0}) and output current (I_d).

The usefulness of this table can best be illustrated by means of an example.

Problem. To specify (*a*) the transformer-secondary voltage rating; (*b*) the transformer-secondary and primary kva ratings; (*c*) the average, rms, and peak current which the rectifying elements will need to carry; and (*d*) the peak inverse voltage which will exist across the rectifying elements, for a diametric (full-wave) rectifier to deliver 10 amp of direct current at 250 volts. Assume that there is no voltage drop in the rectifying elements; also assume a resistance load.

Solution: (*a*) Transformer-secondary voltage each side of center tap $= 1.11 \times 250E_{d0}$
$= 277.5$ volts

Total secondary volts $= 555.0$

(*b*) Transformer secondary kva rating $= 1.75 \times$ d-c, kw

$$= \frac{1.75 \times 10 \text{ amp} \times 250 \text{ volts}}{1000}$$

$= 4.37$ kva

Transformer-primary kva rating $= 1.235 \times$ d-c kw

$$= \frac{1.235 \times 10 \times 250}{1000}$$

$= 3.09$ kva

(*c*) Rectifying-element currents

Average $= 0.5 \times 10I_d = 5$ amp

Rms $= 0.786 \times 10I_d = 7.86$ amp

Peak $= 1.57 \times 10I_d = 15.7$ amp

(*d*) Rectifying-element peak inverse voltage $= 3.14 \times 250E_{d0} = 785$ volts. (The actual ratings will need to be increased to compensate for the voltage drop in the rectifying elements, transformer reactance, and the like.)

2. NON-CONTROLLED RECTIFIERS

The types of non-controlled electronic rectifying devices in common use are: (1) high-vacuum tubes; (2) hot-cathode gaseous rectifier tubes; (3) metallic rectifiers.

Non-controlled rectifiers are used in applications where a fixed amount of d-c voltage is required. Applications for non-controlled rectifiers include battery charging, motor-field excitation, generator-field excitation, magnetic-chuck excitation, lifting-magnet excitation, d-c control power supplies, etc.

The high-vacuum or kenotron rectifier tubes are inherently low-current tubes, because the voltage drop and therefore the power loss in the tube are proportional to the current flowing. High-vacuum rectifier tubes may be classified as low or high voltage. Low-voltage types are used primarily for small amounts of d-c control power for other electronic equipments. Often low-voltage tubes are constructed with two rectifying devices in one tube, thus making a single tube suitable for a diametric (full-wave) rectifier circuit. High-voltage types are used in such applications as dust precipitators and high-voltage power supplies for other electronic equipment.

The outstanding characteristic of hot-cathode gaseous-type rectifier tubes is their inherently low and constant voltage drop which results in high efficiency in high-current applications. These tubes are available in a range of current ratings of 0.1 amp to 20 amp and from about 100 volts to 10,000 volts, peak inverse ratings. Low-voltage types, such as are used in battery chargers, are usually filled with argon. Xenon-filled, argon-and-mercury-vapor-filled, and mercury-vapor-filled tubes are used in applications requiring 750, 2000, and 10,000 volts, peak inverse ratings, respectively. Gaseous-type rectifiers will not operate successfully in parallel without load-balancing devices, because the tube drop of two paralleled tubes may be slightly different and the tube with the lower drop will carry all or most of the load current. A more complete discussion of hot-cathode gaseous-type rectifier tubes of both the non-controlled and the controlled type is presented in Section 4.

Copper oxide and selenium-type metallic rectifiers are used predominantly in the low-voltage high-current field. The general characteristics of these rectifiers are a relatively low peak inverse voltage per rectifying disk and a relatively high current capacity. Stacks consisting of one or more cells can be used in series or in parallel to increase their voltage or current capacity, respectively. The copper oxide rectifier is older, from the standpoint of general usage, than the selenium rectifier. The selenium rectifier can operate at a higher temperature than the copper oxide rectifier, and therefore for a given rating it is a somewhat smaller unit than the copper oxide type. Selenium rectifiers are usually operated nearer their breakdown voltage rating than copper oxide rectifiers; therefore the selenium rectifier is usually more subject to damage on overvoltage. Other materials, such as copper sulfide, exhibit the same rectifying action as the copper oxide and selenium materials but are not as commonly used.

3. CONTROLLED RECTIFIERS AND INVERTERS

Controlled rectifiers generally are used whenever it is desired to adjust the d-c output voltage level over a reasonably wide range, or when it is necessary to regulate the output voltage to compensate for changes in the load current or changes in the input line voltage. Typical applications of controlled rectifiers are: (1) adjustable d-c voltage for motor- and generator-field supplies; (2) adjustable d-c voltage for d-c motor armatures; (3) adjustable d-c voltage for the d-c windings of saturable reactors which, in turn, control motor, lighting, or resistance power circuits; (4) adjustable d-c voltage supply for testing of various d-c devices; (5) d-c power for charging capacitors in energy-storage resistance welders at a given rate and to a given voltage.

In small rectifiers, it may sometimes be more convenient and economical to adjust the rectifier d-c output either by adjusting the a-c voltage input or the d-c voltage output by means of a slide-wire resistor. Also in large rectifiers it may be desirable to adjust the output voltage by means of induction-voltage regulators, adjustable auto transformers, or saturable reactors in the a-c circuit. If these methods are used, the rectifying elements can be of the non-controllable type.

Grid-controlled thyratron rectifiers provide greater flexibility, faster response, and less bulky control equipment than rectifiers controlled in the power circuits. Furthermore, automatic control in larger sizes is generally more economical with thyratron control than with power circuit control.

Thyratron-type rectifiers are always built in a single envelope. Some thyratron tubes are controlled electromagnetically by a plate on the outside of the tube. By far the most common practice, however, is to control the tubes electrostatically with a grid in the electron path. Thyratron tubes can be obtained in ratings as low as approximately 20 ma and have been built in ratings as high as 100 amp. (See Section 4.) The maximum size of standard tubes, however, is about 12.5 amp. (High-vacuum triodes are generally not used in controlled rectifier circuits.)

Controlled rectifier tubes may be connected in any of the circuits shown in Table 1. Tubes and transformers, however, should be carefully selected so that the ratings are not exceeded. As mentioned earlier, the peak and rms currents in a rectifier will be higher on resistance and counter-emf loads than on inductance loads. Not only must this factor be considered, but also consideration must be given to the higher peak and rms currents that will result from the use of grid control on loads such as d-c motor armature circuits. Another factor to be considered is the unbalance in tube currents which may exist in polyphase rectifiers operating with rated current and having a large amount of phase retard. Figure 1 shows the voltage and current wave shapes of a diametric (full-wave) rectifier operating with varying degrees of phase retard on an inductive load, a resistive load, and a fixed counter-emf load all having different electrical characteristics. Although a diametric rectifier is used here for the sake of clarity, the same fundamental information applies to other rectifier circuits. Although three different load classifications are shown, many loads consist of various combinations of resistance, inductance, and counter emf.

The load current on a very highly inductive load is nearly constant even with a diametric rectifier circuit. As the firing angle is retarded, the energy from the inductance is transferred back into the line circuit during a portion of the cycle when the tubes would be normally non-conducting. When the firing angle has been retarded approximately 90°, the inductance current is theoretically zero. Practically, however, the 90° firing point will result in a current of approximately 10 to 20 per cent of the maximum which would be expected as calculated by Ohm's law.

In a sense a purely resistance load is somewhat academic because this type of load seldom occurs in practice. It is, however, a logical stepping stone in estimating what is to be expected from counter-emf loads. It can be observed from Fig. 1 that the load current in a resistance load follows identically the wave shape of the output voltage.

Counter-emf loads fall into two general classifications: fixed and variable. A battery is a typical example of a fixed counter-emf load. The peak value of load current in a fixed counter-emf load is determined by the difference between the peak a-c voltage and the counter-emf potential, and the circuit resistance. Exceedingly high current can be obtained on low-impedance circuits with a very small difference in voltage between the load

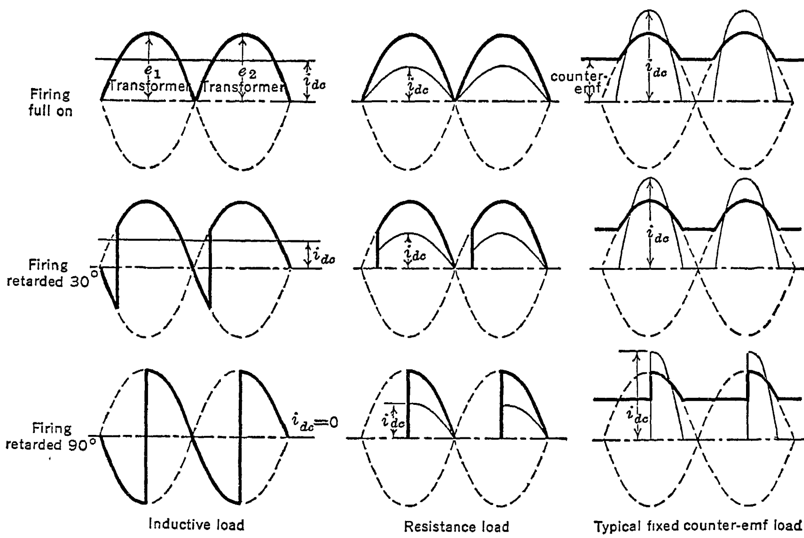


FIG. 1. Rectifier Output Voltage and Load-current Wave Forms for Different Types of Loads and with Several Degrees of Phase Retard with Diametric (Full-wave) Rectifier

and the peak rectifier output. The inherent circuit impedance may reduce the peak current to a reasonable value; if not, external resistance or inductance can be added for that purpose. Figure 1 shows that as the firing angle of a rectifier on a fixed counter-emf load is retarded there is, up to a certain point, no change in the output current, but, beyond the point where the transformer-secondary voltage and the counter-emf voltage intersect, a reduction in current is obtained by a further retard in firing position. A d-c motor armature is a typical variable counter-emf load. This type of load is quite different in its operation from a rectifier or other fixed counter-emf load. A motor armature is a good example of a load which includes resistance, inductance, and variable counter emf.

The thyratron rectifier circuits which have been described can also be used for inverters to convert d-c power into a-c power, provided that both d-c and a-c power supplies are available for the transfer of power and that the grid-control circuits are properly arranged. Many motor-control rectifiers are operated as inverters during reversal of motor armatures and during fast decay of the stored energy in generator fields. In these applications the conventional rectifier circuit is used and the grid firing point is retarded to a point late in the positive half cycle. In the high-power field, highly specialized inverters have been built for several applications.

IGNITRON RECTIFIERS. Ignitron rectifiers are ideally suited to their use in the high-power field, because their cathodes—a mercury pool—can supply electron emission for tremendous overloads. The maximum current is limited by the mechanical forces involved and by the ability of the tube to deionize rapidly enough to prevent arcback. Ignitron rectifying devices are made in capacities from about 12 amp up to 1000 amp d-c continuous rating. (In 1945 approximately 10 per cent of the central-station power generated in the United States passed through ignitron rectifiers.) Continuously evacuated ignitron rectifiers are available in sizes as large as 6000 kw at 600 volts direct.

Figure 2 shows the efficiencies of ignitron rectifiers as compared with synchronous converters and synchronous motor-generator sets. The rectifier efficiency increases with higher voltage because the voltage drop across the arc in the ignitron is nearly the same in all cases, and, therefore, with lower d-c voltage the arc drop has a greater proportional effect than at high voltage.

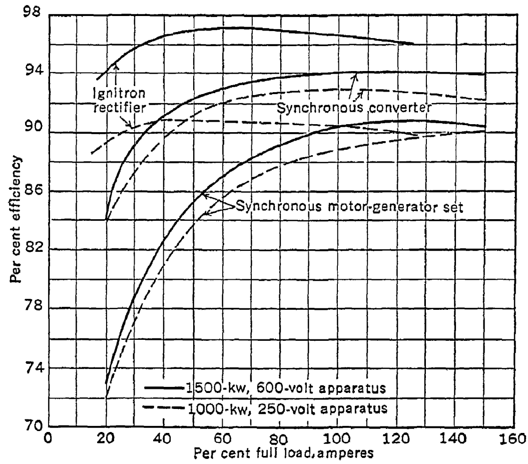


FIG. 2. Comparative Efficiencies of Power-conversion Units

Thyratron rectifiers, like vacuum tubes, may be controlled with very minute power levels on the grid. An ignitron, however, requires a considerable amount of power for a short time for firing. This power may be supplied either by magnetic excitation circuits involving no tubes and purely static control devices, or they may be controlled by electronic firing circuits involving thyratrons in the ignitor circuit. The magnetic firing circuit is used where the maximum of reliability is required. Thyratron-type firing circuits, however, may be less expensive to build, and also their inherently fast operating speed may be more desirable for some applications.

4. THYRATRON AND IGNITRON CONTACTORS

Two controlled rectifier tubes of the thyratron or ignitron type can be connected in inverse parallel and the combination connected in series with a single-phase load to make a single-pole single-throw a-c switch called a thyratron contactor or an ignitron contactor. Likewise, several pairs of tubes may be connected in polyphase circuits to make a polyphase single-throw a-c contactor. This type of contactor has several salient features:

1. In applications involving a very large number of operations, it eliminates the mechanical wear and resultant maintenance of mechanical-type contactors. Furthermore, it is quiet in operation.
2. This type of circuit is inherently fast in response. This means that power circuits may be closed or opened more rapidly than with conventional magnetic contactors.
3. When these contactors are used with the proper phase control systems, they may be fired synchronously at a given phase position in each half cycle to avoid the transient currents that will result in inductive loads if the power circuit is closed at random.
4. These contactors may be used to control the effective a-c load voltage by adjustment of the firing point. Control of load power can be obtained without power loss and without undue voltage regulation due to load changes.

The choice of thyratron or ignitron tubes is dependent on the magnitude of power involved. Typical applications of thyratron-type contactors are: (1) for controlling the speed of some types of a-c motors; (2) for controlling the output voltage of high-voltage transformers; (3) for low-power resistance welders; (4) for high-voltage resistance welders. By far the largest number of applications of electronic contactors, however, has been made in the resistance welding field, in which ignitron tubes are used.

General requirements for resistance welding are: (1) single-phase power; (2) power impulses having a high peak value over a short period of time; (3) a large total number of impulses over a given period of time; (4) control of the effective amount of welding current; (5) controlled firing to eliminate transient currents. These requirements have made the ignitron contactor with its extremely high peak-current capacity and its controllability ideally suited to resistance-welding applications.

Figure 3 illustrates a typical single-phase ignitron contactor for resistance-welding service. This contactor is of the simple on-off type. It is used to give faster operation than magnetic-type contactors as well as to eliminate the maintenance on magnetically operated mechanical contactors. Figure 4 shows an elementary circuit diagram for an on-off type ignitron contactor. This circuit operates as follows: When the initiating switch is open, no current will flow because power is not being supplied to the ignitors for firing the tubes. Assume now that the initiating switch is closed and that anode *a* is positive. Current will not flow through the left-hand tube until the ignitor has been fired. The ignitor firing circuit starts with anode *a*, continues through the lower right-hand metallic rectifier, through the water-flow switch, fuse, and initiating switch, and through the upper left-hand metallic rectifier into the ignitor which is immersed in the mercury pool. Current flows from this pool to the primary of the welding transformer and back to the other side of the a-c line. This means that the entire line voltage and the entire load current are available for firing the ignitor. With this voltage and current available, the ignitor fires the ignitron tube, and then load current flows through the main anode of the left-hand ignitron. When the anode *b* is positive, ignitor firing current flows through the lower left-hand metallic rectifier, through the initiating switch, fuse, and water-flow switch; then through the upper right-hand metallic rectifier into the mercury pool and back to the other side of the a-c supply. The metallic rectifiers provide a path for the ignitor firing current during the half cycle in which a specific ignitor is to be fired. During the other half cycle, the metallic rectifiers prevent the flow of reverse ignitor current which would damage the ignitor.

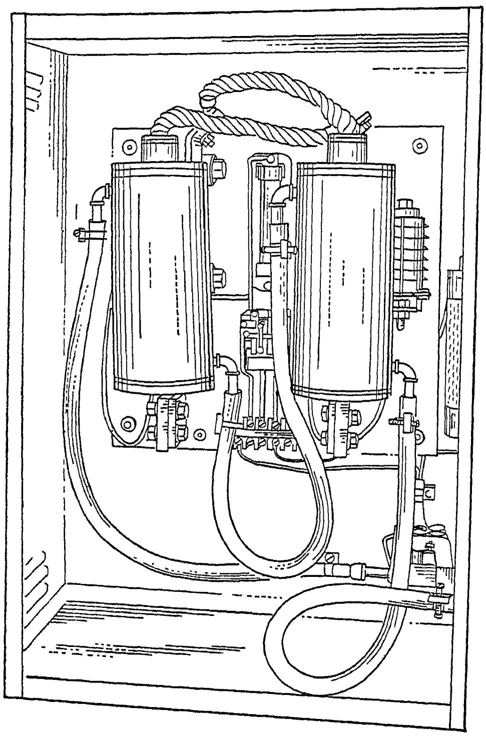


FIG. 3. Single-phase Ignitron Contactor with Size D Tubes (Courtesy General Electric Co.)

Figure 5 shows how thyratron tubes may be used to control the ignitron power tubes in order to obtain phase control or synchronous timing control of the output power. A resistor is used in the anode of the two thyratron tubes to limit the peak current flowing while a fuse is used to protect the tubes against high rms current which would result if the ignitor were continuously fired with no load current flowing.

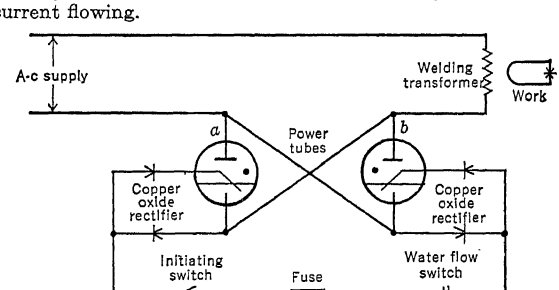


FIG. 4. Circuit Diagram for an Ignitron Contactor with Metallic Rectifier Firing

Figure 6 is a typical duty cycle rating curve for four sizes of contactors (ratings are for two tubes) used on a 460-volt power supply. Three factors affect the amount of current which can be handled by a pair of ignitrons: (1) The maximum current is a function of the line voltage; for low line voltages, the tubes will carry a higher current than for high line voltages. (2) For a given line voltage there is a maximum rms current rating irrespective of the length of time the current flows. (3) If the current flows over an appreciable length of time, even though the duty cycle is low, the averaging time of the tube must be considered to prevent overloading the tube from a thermal standpoint. Per cent duty shown in Fig. 6 indicates the percentage of total time that current is flowing through the tubes.

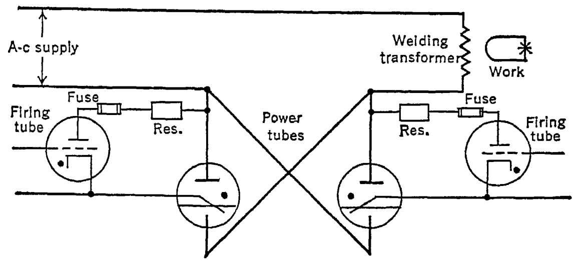


FIG. 5. Ignitron Contactor with Thyratron Firing Circuit

A pair of size D tubes has a continuous rating of 800 amp. As the duty cycle is reduced, however, the maximum current can be increased. If the time involved is short, the rating can reach almost 5000 amp. The size D tubes have an averaging time of 5.6 sec. This means that, for any 5.6-sec period, the average current should not exceed 800 amp rms, even though the actual current during conduction equaled 5000 amp. In other words, the tubes could carry 5000 amp for approximately 0.9 sec (800/5000 × 5.6), provided no

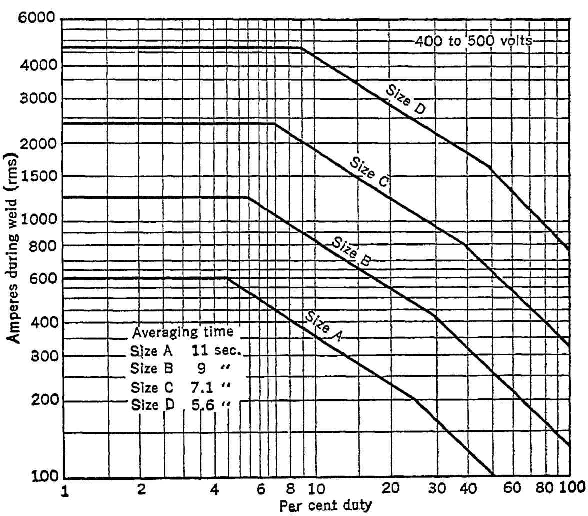


FIG. 6. Ignitron-contactor Duty-cycle Rating Curve

current was carried during the remainder of the 5.6-sec period. If it were necessary to have current flowing over a 5.6-sec period continuously, the tube rating would be the same as for 100 per cent duty, or 800 amp.

FUNDAMENTAL ELECTRONIC CONTROL CIRCUITS

Because the applications of electronic control are so diversified and because most applications are highly specialized it is impractical to describe even a moderately complex circuit with all its ramifications. An attempt will be made in this section to discuss several fundamental control circuits which are commonly used in complete electronic systems, with the hope that these circuits will be recognized when they are a part of complete systems.

5. STABILIZED D-C CONTROL POWER SUPPLIES

Many electronic control equipments include a small d-c power supply usually consisting of a diametric (full-wave) rectifier and a suitable filter. These are similar to the power supplies used in radio receivers which are discussed in Section 7.

In order to provide stability of operation of the associated electronic control equipment, it is often necessary to provide a d-c output voltage which is held constant irrespective of line voltage or changes in load. This may be accomplished in several ways; the most common are:

1. A-c voltage stabilizers. Figure 1 illustrates one typical type of automatic voltage stabilizer circuit. The voltage-regulating action is due primarily to reactor 1 and the parallel capacitor. When the line voltage is high, reactor 1, which operates near the knee of the *B-H* curve, becomes saturated. Under this condition the reactor current and the capacitor current are about equal, the total current being at about unity power factor. As a result there is a voltage drop across reactor 2. When the line voltage drops,

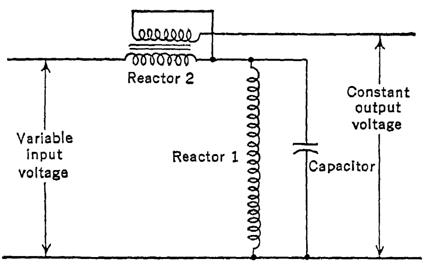


FIG. 1. Circuit Diagram of A-c Voltage Stabilizer

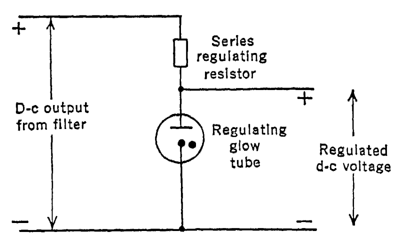


FIG. 2. Glow-tube-type D-c Voltage Regulator

however, the reactor current falls off more rapidly than the capacitor current. As a result the current through reactor 2 is predominantly capacitive and there is a rise in voltage across reactor 2. Units of this type will hold the output voltage to within ± 1 per cent over about a ± 15 per cent change in input voltage.

2. Gaseous-type voltage-regulating glow tubes. These tubes have inherently a constant voltage drop. When the voltage has reached the ionization point a very small increase in applied voltage will result in a comparatively large increase in tube current. These tubes are well suited to control power circuits involving small currents. Three ratings are most commonly used, namely, 75, 105, and 150 volts at 40 ma maximum. Figure 2 shows a regulating circuit, using a glow tube. The constant-voltage characteristic of the glow tube will permit the rectifier output voltage to be increased or decreased, and the voltage difference will be largely absorbed by the series regulating resistor. This results in a nearly constant output voltage across the glow tube.

3. A series-type voltage regulator. This regulator uses a glow tube merely as a reference voltage and involves a regulating principle which will be discussed later. This type of voltage-regulating system will provide larger amounts of voltage and current and at greater accuracy than can generally be obtained with the simple glow-tube arrangement discussed before.

6. TIMING CIRCUITS

Timing circuits are widely used in electronic control systems. In many equipments such as general-purpose timers, resistance-welder sequence timers, timers for cut-off register applications which operate correcting devices for a given length of time, and timing circuits which control the acceleration or deceleration rate of motors, timing is a definite function. Often, however, the use of timing circuits may be incidental to the function to be performed by a given equipment. For example, filter capacitors in d-c control power supplies, capacitors across relay coils which are operated by half-wave rectifiers, resistor-capacitor combinations, and electronic relays involving impulses of very short duration which are used to maintain an operating signal long enough to operate a magnetic relay are all energy-storage-type timing circuits.

A capacitor is used as a basic timing element in most timing circuits. Figure 3*A* shows a simple series resistor-capacitor circuit. When the switch is closed, the capacitor charges through the series resistor. The instantaneous voltage across the capacitor is expressed by the equation $e_{ct} = E(1 - e^{-t/rc})$, where t = time in seconds after the switch is closed. Figure 3*B* shows a similar circuit in which the capacitor is charged almost instantly (assuming that the d-c source has no resistance) and in which the capacitor discharges through the parallel resistor. The instantaneous capacitor voltage after opening the switch is expressed by the equation $e_{ct} = E \times e^{-t/rc}$. The above equations will give the voltage at any time. A more generally used term, however, is the time constant. The time constant is defined as the time at which the capacitor has charged to approximately two-thirds of its final voltage value or the time at which the capacitor has discharged to approximately one-third of its initial voltage value. The time constant is expressed as $T = rC$, where T = time in seconds, r = resistance in ohms, and C = capacitance in farads. Often a basic timing circuit will include a parallel r–C circuit with a resistor in series. This will result in a time lag in the charging circuit and a lag in discharging as well.

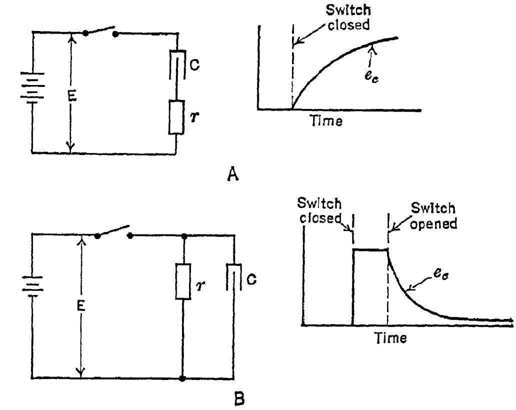


FIG. 3. Basic *R-C* Timing Circuits

r-C circuits are used in complete circuits to obtain a time delay by applying the capacitor voltage to a tube grid to render the tube conducting or non-conducting at a given capacitor voltage. Although different detail circuit arrangements are used in timing, the basic circuits shown are common to most electronic timers.

A general-purpose time-delay-relay circuit which uses a parallel r-C circuit for timing is shown in Fig. 4. This is typical of complete timing circuits. The coil of the relay is energized a definite time after switch S is closed. When switch S is open, the cathode of the vacuum tube is connected through $2R$ (which is a relatively low resistance) to the anode. Resistor $1R$ and the potentiometer P form a voltage divider across the a-c power supply. With the switch open, the cathode is effectively connected to line 1 (through resistor $2R$). Therefore current will flow from line 3 through a section of the potentiometer, through the parallel resistor-capacitor combination into the grid, and back through the cathode. The grid in this instance acts as an anode since it is positive with respect to the cathode during every other half cycle. The voltage drop across the timing resistor R causes capacitor C to become charged. As capacitor C charges, the voltage applied to the grid during alternate half cycles becomes less positive—the grid voltage is the algebraic sum of the a-c voltage and the capacitor voltage. The capacitor will continue to charge until the voltage across the capacitor is equal to the peak of the a-c voltage from line 1 to the slider of the potentiometer. If the potentiometer P is turned to the extreme counterclockwise position, this will be the peak line voltage. As potentiometer P is turned clockwise, the capacitor charges to a lower value. When switch S is closed the cathode is connected to the other side of the line. At this instant the grid of the tube is negative with respect to the cathode by whatever potential the capacitor C is charged. The instantaneous grid voltage, however, is a summation of the d-c capacitor voltage and the a-c voltage from line 3 to the potentiometer slider. Figure 5 shows the action of the capacitor discharge circuit plus the a-c component. It can be seen that, when the switch is first closed, the grid is sufficiently negative at all times so that no plate current flows. As the capacitor discharges, however, the grid potential reaches a point where sufficient plate current flows to energize the relay in the plate circuit. The capacitor C_1 of Fig. 4 across the relay coil is

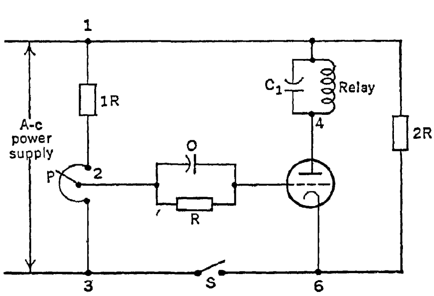


FIG. 4. General-purpose Time-delay-relay Circuit

used as an energy-storage device during the half cycle in which the tube is conducting so that it can supply energy to the coil during the half cycle when the tube is not conducting, thus preventing the relay from chattering.

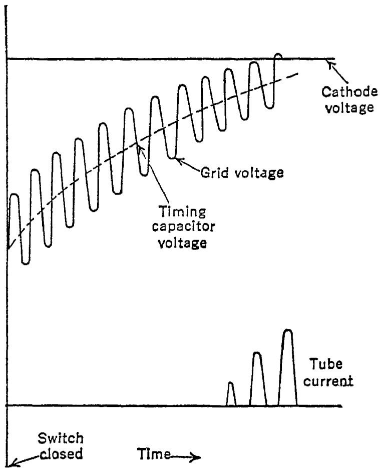


FIG. 5. Grid Voltage and Anode Current after Timing Switch of Timer (Fig. 4) is Closed

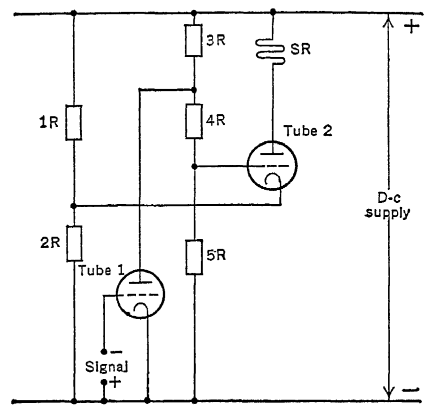


FIG. 6. Typical D-c Voltage Amplifier

7. D-C AMPLIFIERS

D-c amplifiers are commonly used in photoelectric controls, motor and generator controls, electronic voltage regulators, regulated battery chargers, and other equipment where it is desirable to amplify d-c signals. These amplifiers are generally not too well suited for applications where the signal voltage is low and where a high degree of amplification is needed, because the instability of the tube characteristics can result in output variations equal to or greater in magnitude than would be obtained from a complete input-signal range. D-c amplifiers are very well suited, however, to regulating circuits where the difference between two fairly large voltages is applied to the grid circuit. Regulating circuits will be discussed in more detail later.

Figure 6 shows a typical d-c voltage amplifier. Resistors $1R$ and $2R$ form a voltage divider which establishes the cathode potential of tube 2. Resistors $3R$, $4R$, and $5R$ are chosen so that, when plate current is not flowing in tube 1, the grid of tube 2 is positive. When current flows through tube 1, the additional IR drop in resistor $3R$ lowers the grid voltage of tube 2 and reduces its plate current. In the plate circuit of tube 2 is shown a saturable reactor which could be used in a thyratron phase-shifting circuit, although any voltage- or current-responsive device could also be used here. If the signal voltage were zero, the grid of tube 1 would be at cathode potential and tube 1 would be carrying approximately maximum current. As the grid voltage of tube 1 is made negative, the plate current of tube 1 would decrease, thus increasing the plate current in tube 2.

8. A-C AMPLIFIERS

In control circuits a-c amplifiers are usually of the capacitance-coupled type. Though in some circuits they are used in much the same way as in radio circuits they are often employed in circuits in which it is desirable to amplify an impulse signal. For example, in photoelectric cutoff or web-register control systems used in cutting bags or labels at a particular point with respect to the printed material it is desirable for the equipment to respond to small marks on paper which is moving at high speed. The light impulse is obtained by scanning the paper surface with a photo tube. The normal light level may be relatively high, and the mark on the paper may correspond to only a small change in light. Therefore, it is necessary to have a very sensitive amplifier which responds only to rapid light impulses or changes and not to steady-state light or general changes in light level.

A typical impulse amplifier is shown in Fig. 7. Resistor 1*R*, 2*R*, 3*R*, and 4*R* constitute a voltage divider to supply the various voltage levels required in the circuit. The proper grid bias on tubes 1 and 2 is supplied through grid resistors 6*R* and 8*R* respectively. Assume that a steady level of light is applied to the phototube. The circuit constants are arranged so that under this condition current flows in tube 1 but not in tube 2. Now if light suddenly is diminished the phototube current is reduced and capacitor 1*C* momentarily "pulls" the grid of tube 1 to a more negative position. This immediately reduces the plate current in tube 1, which in turn raises momentarily the grid voltage of tube 2 through capacitor 2*C*. Tube 2 is a thyratron type and has a d-c anode supply voltage. Therefore although its grid voltage may go in the positive direction only momentarily this is sufficient to fire the tube and once fired it remains conducting until the reset contact shown is opened.

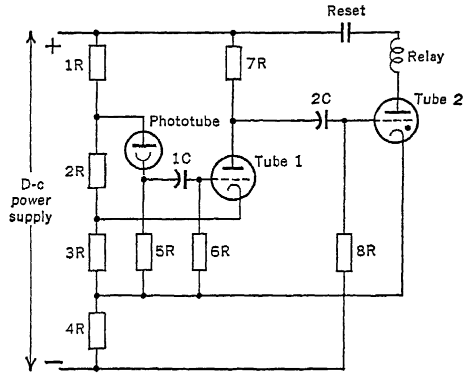


Fig. 7. Typical Impulse-type A-c Amplifier as Used in Photoelectric Cutoff Register Controls

9. REGULATING CIRCUITS

Regulating circuits are used in many applications where it is desirable to hold quantities such as voltage, current, speed, position, pressure, temperature, and the like constant irrespective of conditions, such as load or line-voltage variations, which would normally cause the quantity being held to deviate from the desired value. Any regulating circuit has four fundamental requirements:

1. A standard or a reference voltage, which is held constant at all times and which represents a value against which the quantity to be held constant is compared. The reference voltage is the heart of a regulating circuit because the accuracy of the regulator can never be more accurate than the reference voltage.
2. Means must be provided so that the quantity to be regulated will produce a signal voltage which can be compared with the reference voltage.
3. An amplifier, to amplify any difference that may exist between the regulated quantity and the reference voltage.
4. A controlling means, which will operate from the amplifier to restore the quantity to approximately the value originally held before the deviation occurred.

Figure 8 shows an electronic voltage regulator, the purpose of which is to supply a constant d-c voltage output over a wide range of load current and d-c input line voltage. The total output power from this circuit is limited by the current that can be handled by the 2A3 tube shown. (Parallel tubes or other tube types will deliver more power.) Assume for the moment that switch *S* is open and also that no current is flowing in the plate of the 6J7 tube. Under these conditions, the grid of the 2A3 tube will be at approximately cathode potential. The effective resistance of the 2A3 tube under this condition is very low. Therefore a high voltage will exist across lines 2 and 3. Resistors 1*R* and 2*R*, which are in series with the OA3/VR75 glow tube, are such that with rated regulated d-c output voltage the glow tube will ionize and, once ionized, will conduct a small amount of current. The characteristic of the glow tube is such that, if the voltage between lines 2 and 3 is varied, the voltage across the glow tube will remain reasonably constant. However, the more constant the voltage across lines 2 and 3 can be held, the less variation there will be in the voltage across the glow tube because the voltage across the glow tube changes somewhat with current. Since the objective of this equipment is to hold the voltage output constant, the reference voltage must be held as constant as possible.

The cathode of the 6J7 amplifier tube is connected to the glow-tube anode, point 4. Resistors 3*R* and 4*R* constitute a voltage divider across the output of the regulator and therefore provide a signal to the grid of the 6J7 tube which is proportional to the output voltage. Resistors 3*R* and 4*R* are selected so that with rated regulated d-c output voltage the grid of the 6J7 tube will be at approximately cutoff. The actual grid signal consists of the difference between the voltage across resistor 4*R* and the voltage across the glow tube. If it is assumed that the reference voltage from the glow tube is constant, then any deviation in output voltage will be reflected as a difference in the potential between the grid and the cathode of the 6J7 tube.

As was stated before, if no current is flowing in the 6J7 tube, the effective resistance of the 2A3 tube is very low. The output voltage between lines 2 and 3 therefore will be

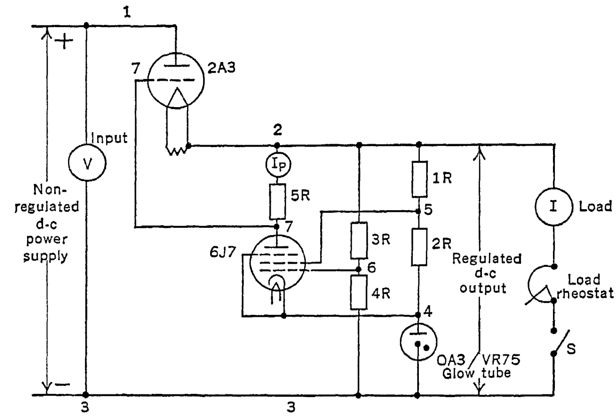


FIG. 8. Electronic D-c Voltage Regulator

high. This, however, results in a grid-to-cathode voltage on the 6J7 tube of a sufficiently high positive value to cause the 6J7 to conduct plate current. This current causes an IR drop between lines 2 and 7 across resistor $5R$, and causes the grid of the 2A3 tube to assume a negative potential. This negative potential increases the effective resistance of the 2A3 tube and thus reduces the output voltage between lines 2 and 3. There is a certain grid-to-cathode voltage on the 2A3 tube which will result in an output voltage across lines 2 and 3 sufficient to adjust the grid potential of the 6J7 tube to give the desired grid-to-cathode voltage on the 2A3 tube.

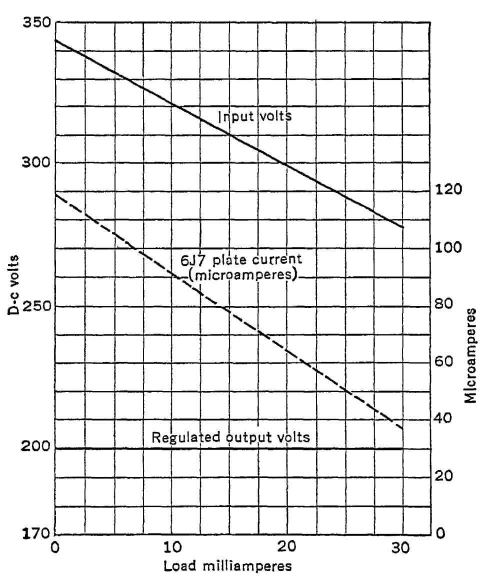


FIG. 9. Performance of Electronic Voltage Regulator Shown in Fig. 8

If now the regulator is holding a given voltage and load is added by closing the switch S, the output voltage between lines 2 and 3 will immediately drop because the effective resistance of the 2A3 tube is in series with the load. In order to correct this low-voltage condition, the regulator must again go into action. When the output voltage is reduced, the grid voltage is made more negative with respect to the cathode voltage on the 6J7 tube. This more negative value of grid voltage reduces the 6J7 plate current and thereby makes the 2A3 grid less negative, which reduces the equivalent resistance of the 2A3 tube, again restoring the correct output voltage. The curves of Fig. 9 show input and output voltage as a function of load current for the regulator. It also shows the plate-current change required of the 6J7 tube to give accurate regulation of the output voltage. It can be seen that, for a wide change in output load current and for a wide change in input line voltage, the regulated output voltage is held nearly constant.

The electronic voltage regulator just discussed is a simple unit when compared with most complex electromechanical regulators; however, it is typical of all regulating circuits in that it includes the fundamental elements of a regulating circuit.

The essential regulating elements can be explained as follows: Referring to Fig. 8, the voltage across the glow tube between lines 3 and 4 is the standard or reference voltage. The voltage across lines 3 and 6 is the signal voltage. Any difference between the voltages 4 to 3 and 6 to 3 is applied to the grid of the 6J7 tube, which amplifies the existing deviation. This in turn controls the grid on the 2A3 tube to correct for the voltage deviation. An important feature of a regulating circuit, as far as obtaining accuracy is concerned, is the use of two relatively large voltages in the comparison circuit. This can be shown best by an example. If the 6J7 tube required 0.1-volt change on the grid to effect a complete swing of the grid of the 2A3 tube, this 0.1-volt change can be accomplished on a 200-volt output circuit with a 75-volt reference voltage by approximately a 0.3-volt total error in the output. In other words, the inherent regulating error of the system does not exceed 0.3 volt out of 200 volts or 0.15 per cent for a wide load and input line-voltage change. Furthermore, if the inherent tube characteristics varied in such a manner that a given 6J7 plate current resulted with a grid voltage 0.3 volt different from the original voltage the output voltage would again be in error by only approximately 0.15 per cent. If lower voltages were used, both errors would increase.

10. THYRATRON GRID-CONTROL CIRCUITS

Thyratron tubes are used in circuits having d-c or a-c anode supplies. When they are used with d-c supplies they are on-off devices and a simple d-c grid-bias control can be used to fire the tubes. In a-c circuits, such as rectifiers or electronic contactors, however, it is usually desirable to control the effective output voltage by controlling the point in each cycle at which a particular tube fires.

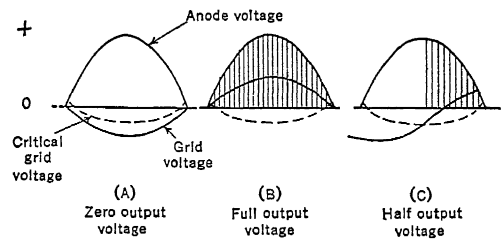


FIG. 10. Method of Grid Control of Thyratron Tubes

When a thyratron is connected to an a-c power supply the anode voltage is different at each point in the cycle. Therefore at the beginning of a positive-voltage half cycle a positive value of grid voltage is required to fire the tube. As the anode voltage increases in the positive direction, the grid voltage required to fire the tube becomes negative. Figure 10 shows a curve of the grid voltage—the critical grid-voltage curve—required to fire a thyratron tube on an a-c power supply. Figure 10*A* shows that with an a-c grid voltage applied 180° out of phase with the anode voltage the tube does not conduct. Figure 10*B* shows that with an a-c grid voltage applied in phase with the anode voltage the tube conducts for the entire half cycle. Figure 10*C* shows that with the a-c grid voltage lagging the applied voltage by 90° the tube conducts for half of the half cycle.

Although a number of different types of grid-controlled circuits can be used for controlling the output voltage of thyratron tubes, the most widely used control circuits are: (1) grid-voltage phase shifting; (2) fixed 90° phase shift with adjustable d-c bias.

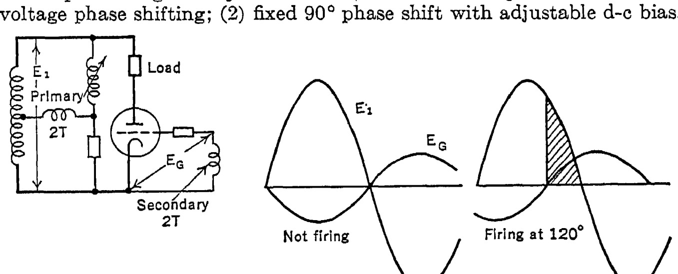


FIG. 11. Basic Circuit for Controlling Thyratron Tubes with Adjustable Grid-voltage Phase Angle

Figure 11 shows the basic circuit for a method of controlling thyratrons by means of an a-c grid potential whose phase angle can be adjusted. The grid-voltage phase position is determined by the ratio between the variable inductance and the fixed resistance shown in the primary of the grid transformer 2*T*. Figure 12 shows the vectors of the voltages involved in the grid circuit. E_1 is the anode-transformer secondary voltage. *Ir* is the voltage across the resistor and is in phase with the resistor-reactor series current. Ix_L is the voltage across the variable reactor. The actual grid voltage E_g can be seen to vary in phase position, but not in magnitude, as Ix_L is changed (change in the reactance). The variable reactor could be an iron-core reactor with a removable plunger, or a small saturable reactor, the d-c winding of which could be energized by a vacuum tube or other means. It is also possible to use a fixed capacitor and a variable resistor to obtain this same type of control. Another method of obtaining this same type of control is by the use of a selsyn (induction phase shifter) which has a three-phase primary stator and a single-phase secondary rotor. As the rotor is turned, the rotor output voltage, which can be applied directly to the grid circuit, changes in phase position.

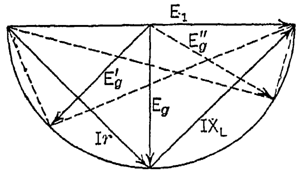


FIG. 12. Vector Voltages of Phase-shift Circuit of Fig. 11

Figure 13 shows the basic circuit for obtaining a fixed a-c voltage phase shift plus an adjustable d-c voltage for controlling the thyratron firing point. The voltage E_g can be seen to be made up of two components: E_2, having a magnitude which is determined by the ratio of transformer 2*T*, and a phase position which is determined by the ratio of resistance and capacitance in the primary of transformer 2*T*; and an adjustable d-c

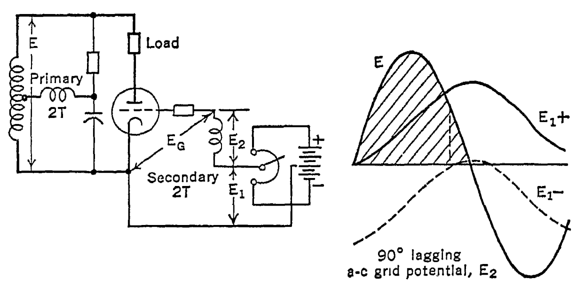


FIG. 13. Basic Circuit for Controlling Thyratron Tubes with Fixed A-c Voltage Phase Shift plus Adjustable D-c Grid Voltage

potential E_1. With this circuit, grid control is obtained by adjusting the d-c grid potential E_1. Figure 13 shows firing full on with positive grid voltage and nearly full off with negative grid voltage. Smooth control can be obtained over the entire range.

COMPLETE ELECTRONIC DEVICES

Many electronic circuits can be built into complete electronic devices, the operation of which is independent of associated equipment. Timing relays, electronically regulated power supplies, rectifiers, and electronic contactors are complete electronic devices which have been discussed previously.

11. ELECTRONIC RELAYS

CONTACT-OPERATED ELECTRONIC RELAYS. Contact-operated electronic relays are used where it is desirable to operate a magnetic contactor or other electrically operated device upon the closure of a circuit which has insufficient current-carrying capacity to operate the final or an intermediate device. For example, it is often desirable to have a power circuit initiated when an instrument pointer reaches a certain mark on the scale. The pointer contacts in most cases are not only inadequate to carry the power in the load circuit but also inadequate to carry the power required to actuate a contactor in the load circuit. The contacts, however, are adequate for insertion in the grid circuit of an electronic relay, which will in turn initiate a magnetic contactor. Other applications of this type of relay are: (1) liquid-level controls where the liquid itself is the conducting medium—in this case a pump or a valve is operated by means of the relay; (2) high-low gaging of small parts where the parts passing between contacts carry the grid signal which in turn operates the relay to actuate reject devices; (3) drop-switch circuits in textile mills to operate signals which indicate broken "ends" of yarn.

Figure 1 shows a typical circuit diagram for a contact-operated electronic relay. When the switch S is open, the transformer secondary supplies power through the half-wave rectifier to charge capacitor $2C$. At the same time the grid of the amplifier tube is at cathode potential. The triode tube conducts full current and relay CR is energized. When the switch S is closed, the grid of the amplifier tube is brought to a negative potential and the relay CR is de-energized. By the proper choice of resistors $1C$, $1R$, and $2R$, this circuit will operate with resistances as high as 10 megohms across the switch terminals. Also, with other resistor and capacitor values it will operate when the switch circuit is closed only long enough to give a very short impulse for charging capacitor $1C$.

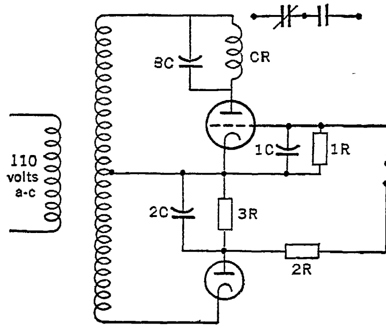


FIG. 1. Typical Circuit Diagram for Contact-operated Electronic Relay

PHOTOELECTRIC RELAYS. Photoelectric relays are often used as limit switches for operating signals, counters, or other devices in applications where mechanical limit switches would be unsuitable because of an excessive number of operations, because of extremely high velocities of moving articles, because of temperature extremes or because the material which is to operate the device has insufficient mechanical strength to operate a mechanical limit switch. There are also many photoelectric-relay applications which are unique to light-responsive devices, as indicated by the following examples: (1) operation by the light reflected by or transmitted through certain colors of material; (2) operation by the amount of light transmitted through holes in cloth, paper, steel, or other material; (3) operation by light reflected from certain types of bottle cracks.

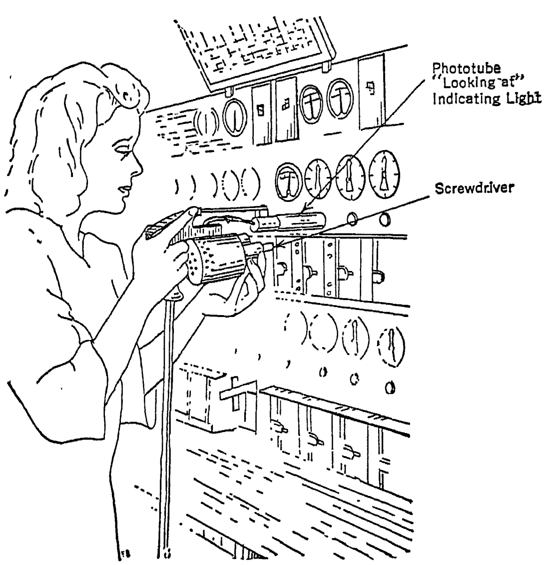


FIG. 2. Application of Photoelectric Relay to Motor-operated Screwdriver (Courtesy General Electric Co.)

Figure 2 illustrates an ingenious application of a simple photoelectric relay. A motor-driven screwdriver is used to make an adjustment on a relay contact. When the contact opens, an indicating light is turned on. A phototube is held over the indicating light to shut off the motor-driven screwdriver automatically when the light comes on, thus assuring proper adjustment of the relay.

A number of special and general-purpose types of photoelectric relays are commercially available. Figure 3 shows several general-purpose photoelectric relays: 1 is an outdoor-type weatherproof relay; 2 is an indoor-type high-sensitivity relay with a large light-collecting lens; 3 is a high-sensitivity relay with a separately mounted phototube; 4 is a simple general-purpose relay; 5 is a general-purpose relay with a separately mounted phototube.

Figure 4 shows the circuit diagram for the general-purpose relay, numbered 4 in Fig. 3. Assume that light is not shining on the phototube and that the slider of potentiometer P

is turned completely counterclockwise. Therefore the full voltage of the lower transformer secondary is connected from the cathode to the grid circuit. During every other half cycle the grid circuit becomes positive. This charges the grid circuit capacitor by grid rectification in the direction indicated. The grid circuit is positive during the half cycle

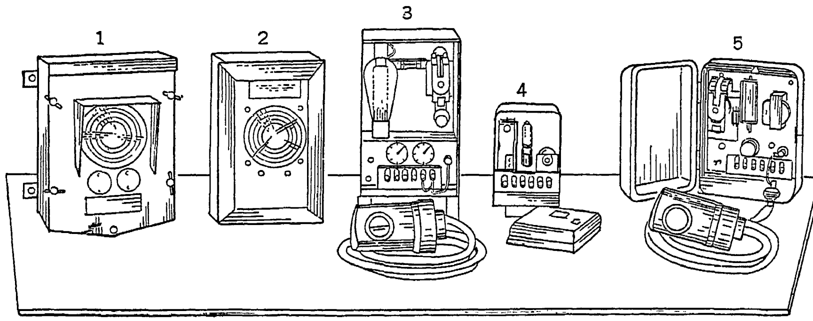


Fig. 3. General-purpose Photoelectric Relays (Courtesy General Electric Co.)

in which the plate is negative. Thus no plate current flows. When the plate circuit is positive, the polarity of point A is negative; also the charge on the capacitor is in a direction to make the grid even more negative, thus preventing the relay from being energized. Now, if light is applied to the phototube, current flows through the phototube and capacitor in the direction of the arrow during the half cycle in which the grid is negative. This effectively discharges the capacitor in proportion to the amount of light on the phototube. If sufficient light is applied to the phototube, the capacitor charge due to the phototube will be greater than the charge due to grid current and the grid will become positive, thus energizing the plate-circuit relay. When the potentiometer slider is in the counterclockwise position, the capacitor charges to a higher value than when it is turned clockwise; therefore more light is required to operate the relay in the counterclockwise position of the slider than in the clockwise position.

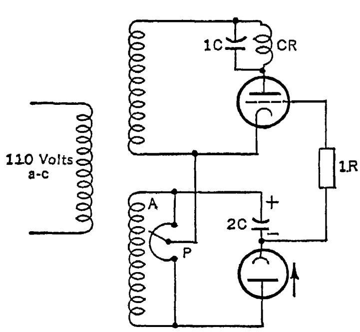


Fig. 4. Circuit Diagram of Relay 4 in Fig. 3

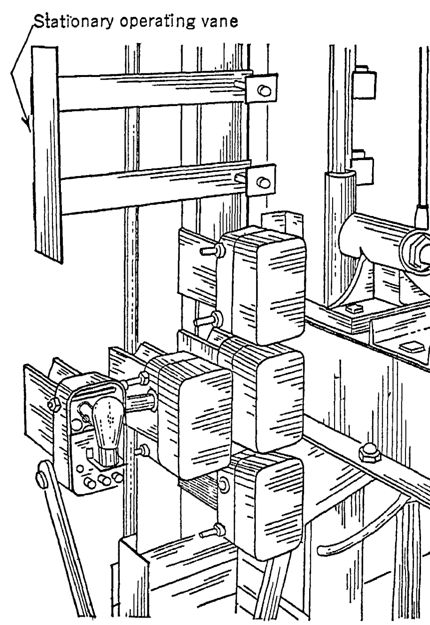


Fig. 5. Oscillator-type Elevator-leveling Relays Mounted on an Elevator Car (Courtesy General Electric Co.)

The a-c operated type of photoelectric relay just described is not suitable for applications where the light impulses are of extremely short duration, because an impulse may occur during the time when the sine wave of anode potential is passing through zero and therefore the relay would not be energized. Impulses of light for a-c operated relays must be at least in excess of 1 cycle. In order to operate relays where the light impulses are of extremely short duration, it is essential to use a d-c power supply for the anode power of the amplifier tube, as well as for the photoelectric tube itself. The circuit diagram of a high-speed photoelectric relay is shown in Fig. 7, p. 21-17.

ELECTROMAGNETICALLY AND ELECTROSTATICALLY OPERATED OSCILLATOR-TYPE RELAYS. The electromagnetically operated oscillator-type relays were probably first used industrially in the leveling of elevators. Figure 5 shows a typical installation of five elevator-leveling relays mounted on an elevator car. Figure 6 shows the connection diagram for this unit. When there is no magnetic material in the oscillator coil circuit, the tube current is low and the relay *CR* is de-energized. As the car approaches a floor, however, an iron channel enters the oscillator-coil magnetic circuit and oscillation stops. This increases the oscillator plate current and energizes the relay *CR* which in turn operates the necessary control equipment to stop the car at the desired level.

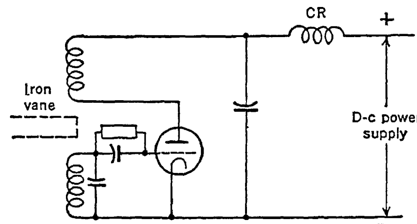


Fig. 6. Circuit of Oscillator-type Elevator-leveling Relay

Figure 7 illustrates the use of an oscillator-type relay which is used in a pyrometer controller. An iron vane on the instrument pointer passes through the oscillator coil and energizes a 15-amp control relay when the pyrometer reaches a predetermined temperature. The point at which the oscillator stops oscillation is very sharp and gives a high degree of control accuracy. Also since no mechanical forces are involved, such as there would be with mechanical contacts, the accuracy of the instrument is not impaired.

Electrostatically operated oscillating-type relays are used as level controllers in grain bins and in tanks which store non-conducting liquids.

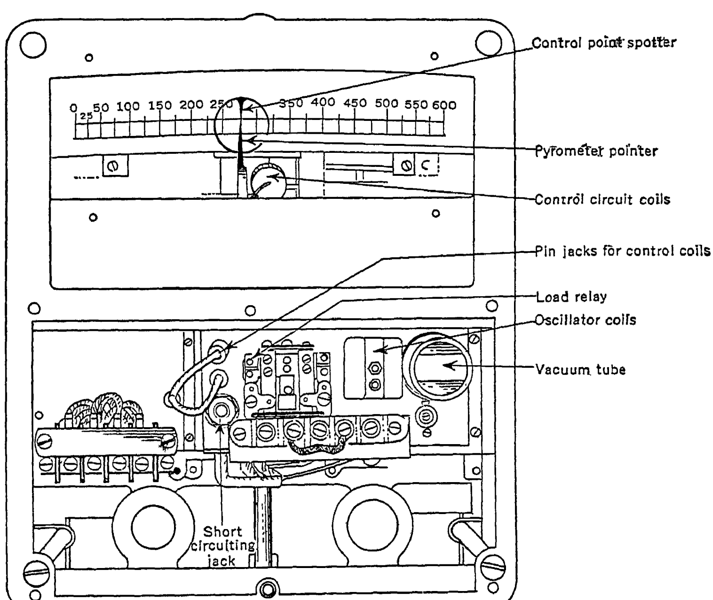


Fig. 7. Free-vane Electronic Pyrometer Controller (Courtesy The Bristol Co.)

12. RESISTANCE-WELDER CONTROLS

The first large-scale applications of electronic control industrially were made in the resistance-welding field. Electronically controlled resistance welders produced results in welding which were previously impossible. Thus a highly specialized branch of control has been developed to meet the many requirements of resistance welders. Resistance welders are divided into three primary types: (1) spot or projection welders; (2) seam or roll-spot welders; (3) upset or flash welders. Most electronic welder-control equipments are used on spot, projection, and seam welders.

The sequence of operation in making a spot weld is as follows: (1) The welder electrodes are applied to the metal during the "squeeze" time. (2) Welding current is passed through the material for a predetermined "weld" time. (3) The electrodes are held closed during a "hold" time, while the material hardens. A pulsation spot weld is made in the same manner as a spot weld except that during the "weld" time the power is intermittently applied with cooling intervals between the power intervals. Seam welds are overlapping or non-overlapping welds made consecutively by intermittent power pulses while the material is passing between two welding wheels.

Most welder controls consist basically of a thyratron or ignitron contactor connected in the primary of a welding transformer. Welding power levels vary from a few hundred volt-amperes for small parts to values in excess of 1000 kva for heavy parts and structures. Because the resistance of the material being welded is very low, a very high secondary current (1000 to 100,000 amp) is required to produce a weld. The relatively long secondary "loop" which is used, in combination with the low resistance of the secondary circuit, results in a low power-factor load.

From a power-system standpoint, a single-phase load with such high peak-power demands of very short duration is sometimes a disadvantage. Energy-storage-type welder controls store energy either in capacitors or in an inductance during the non-welding time and "dump" the stored energy into the material to produce a weld. These controls have three advantages over the conventional single-phase controls, particularly when used in welding aluminum or other high-kva-demand materials: (1) The control operates from a three-phase power supply. (2) The peak-power demand is low. (3) The power factor is higher than for most single-phase controls. Energy-storage type of controls are more expensive and less flexible than single-phase controls, and their general acceptance is thereby limited.

The simplest form of electronic control for resistance welders is the single-phase thyratron or ignitron contactor described previously. This provides high-speed operation and low maintenance, but it is simply an on-off power control. Current starts to flow when the initiating switch is closed, and it stops flowing at the first current zero after the initiating switch is opened. The energy in watt-seconds which is delivered to a weld = I^2RT. The resistance R is dependent on the material being welded. For a given resistance the welding current I is directly proportional to the effective voltage. With an on-off type control the current can be changed only by adjustment of transformer taps. Often manual or mechanical timers are used with ignitron contactors to control the time T.

In some cases the approximate adjustment of current magnitude provided by tap-changing combinations on the welding transformer are sufficient. However, certain metals and alloys which must be welded rapidly (within a narrow temperature range) require more accurate current settings. Phase-shift heat-control equipment can be added to ignitron contactors to provide these accurate settings. Figure 8 shows oscillographs of welding currents under three conditions of heat control.

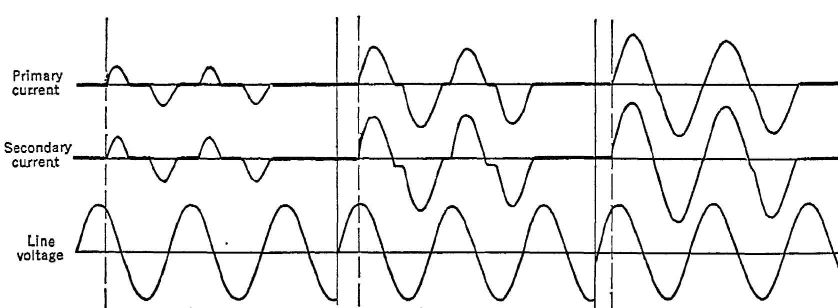


FIG. 8. Oscillograms of Current in Welder when Using Phase-shift Heat Control

Automatic weld timers are often used with electronic contactors to coordinate the various mechanical operations as well as to time the welding power impulse. When welding thin-gage materials, the required time of current flow may be as short as 2 cycles. A variation of 1 cycle would mean a variation in the heat input of 50 per cent and could give a faulty weld. Therefore precision-type timers are required. On the other hand, if the welding time required were 30 cycles a precision timer would not be so important.

If the power tubes are not initiated at a normal current zero a transient current will be present which will result in additional heat in the weld. On short time intervals the transient may add considerably to the total heat. Variations in welds will therefore result from random firing. Synchronous firing is often provided to minimize the transient current by firing the power tubes at a fixed point in the cycle. Figure 9 shows the transient current which will exist when current is started at the zero point on the voltage wave. Figure 10 shows how the transient can be eliminated by starting the current at a normal current zero.

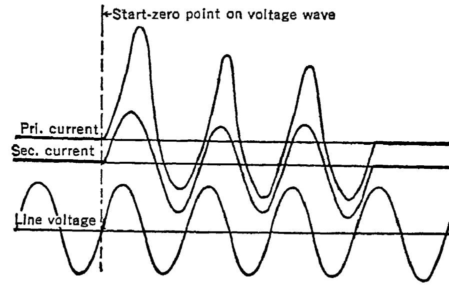


FIG. 9. Oscillogram Made When Current is Started at Zero Point on Voltage Wave

Figure 11*A* illustrates a circuit which will provide phase-shift heat control and synchronous firing. The thyratron firing circuits on both tubes are identical except for the instantaneous polarities of the transformers in the grid circuits. The instantaneous polarities indicated exist when the anode of the left-hand thyratron tube is positive. Figure 11*B* shows the supply voltage E and the grid voltage. The grid voltage consists of two components: (1) E_B, which is the voltage of $3T$ secondary; (2) E_P, which is the voltage of $1T$ secondary. Transformer $1T$ is a "peaking" transformer which has only a small amount of iron in the core. This results in saturation at a particular point in the wave and therefore a peaked output voltage, about 5° or 10° wide. The phase position of the voltage applied to the peaking transformer and therefore the phase position of the voltage peak is adjusted by the phase-shifting rheostat $1R$. The magnitude of E_p when added algebraically to E_b is insufficient to fire the thyratrons. Figure 11*C*, however, shows the grid voltage when switch *SW* is closed. The secondary voltage E_c of $2T$ is of a polarity that subtracts from E_B. Voltage E_p is now of a sufficiently high value to fire the thyratrons. During the half cycles when the grids are positive, the grid capacitors are charged in the direction indicated. This provides a small negative grid bias to prevent misfiring at the beginning of a cycle when the a-c grid voltage is going through zero.

A complete welder control can be made up of a number of individual components, or complete control units can be obtained which include some or all of the above features. Figure 12 is an elementary diagram of a spot-welder control that provides precise timing of the weld impulse, as well as synchronous firing. It does not provide phase-shift heat control. The control essentially is made up of a d-c control rectifier (not shown), a "keying" tube to insure starting the power flow at the desired point in the voltage cycle, a timing circuit which is also a leading firing circuit, a trailing firing-control circuit, and an ignitron power circuit. The keying tube is normally held non-conducting by the negative bias across resistance $2R$. It is fired at a particular point in the wave as determined by the adjustment of resistance $4R$ in series with the primary of the peaking transformer. When the initiating switch $S1$ is in the 1 position, the grid of tube 2 is negative. After the initiating switch is closed in the 2 position, and at the proper time in the next cycle, the grid of tube 2 is made sufficiently positive so that it conducts. This in turn fires the ignitron power tube. As capacitor $2C$ charges (the time being dependent on the setting of resistance $6R$), the grid of tube 2 becomes less positive. At a given time, tube 2 will again have a negative grid voltage and therefore will not fire the ignitron. On an inductive load (a welder load), the current in the left-hand ignitron tube will still be flowing when the voltage applied to the right-hand ignitron tube becomes positive. This means that, as the current in the left-hand ignitron goes to zero, the feedback transformer will put a positive voltage on the grid of tube 3, thus making the right-hand ignitron tube

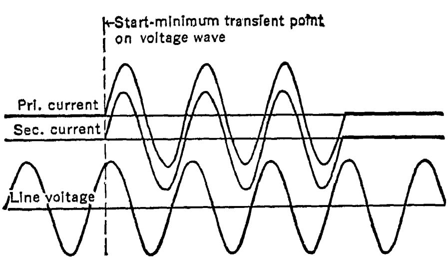


FIG. 10. Oscillogram Made When Current Flow is Started at the Normal Current Zero by Synchronous Firing

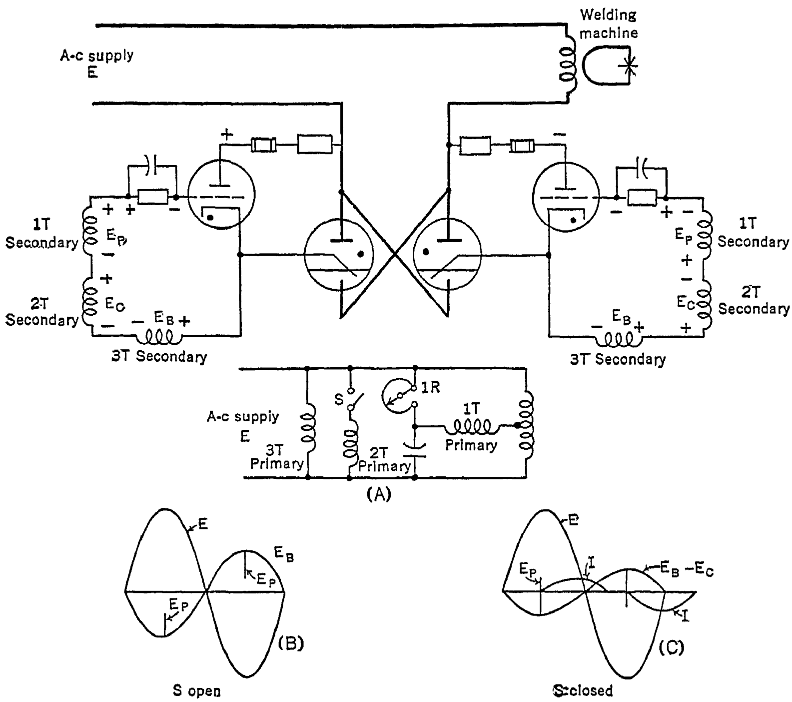


FIG. 11. Phase-shift Heat Control and Synchronous Firing is Provided by the Circuit

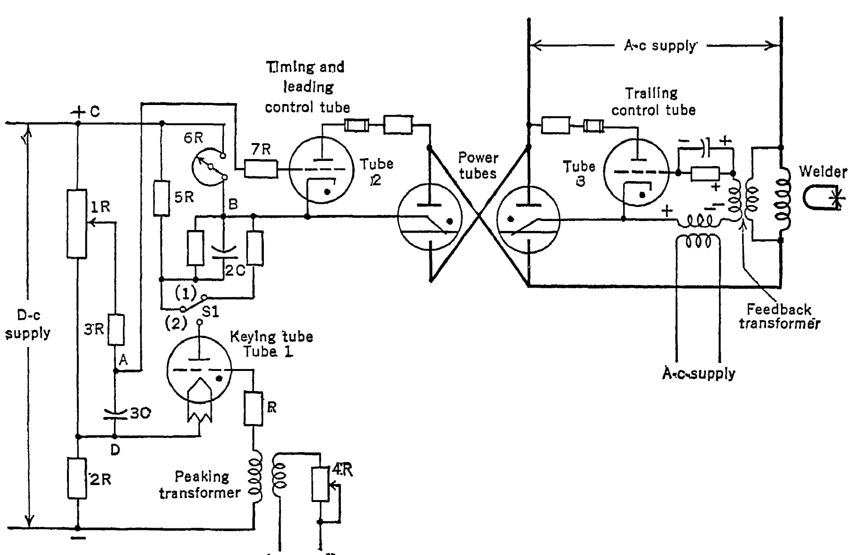


FIG. 12. Circuit of Welder Control Having Precise Timing and Synchronous Firing, but without Heat Control

conduct for a half cycle. For each half cycle which the left-hand ignitron tube fires, there will be duplicate firing on the right-hand ignitron tube. This circuit provides a means of controlling two tubes with different cathode potentials from a single timing source. Furthermore, it eliminates the possibility of obtaining an odd number of half cycles of power, which would cause saturation of the welding transformer. This circuit, however, cannot be used for heat control.

Figure 13 shows a fully electronic welder-control panel. The top section includes an electronic sequence timer for timing the various parts of the machine cycle. The center section includes a synchronous precision weld timer and phase-shift heat control. The bottom section is an ignitron contactor.

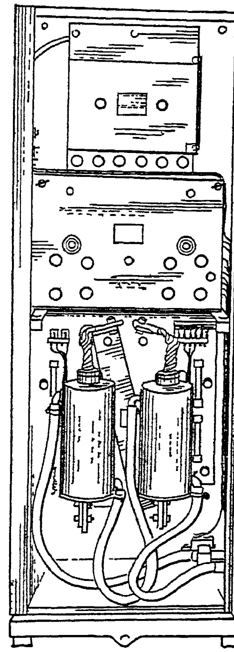


Fig. 13. Fully Electronic Welder Control Having Precise Timing and Synchronous Firing, but without Heat Control (Courtesy General Electric Co.)

13. D-C MOTOR CONTROL

Electronic adjustable-speed drives for d-c motors have been applied in many industries. Typical of these applications are: (1) drives for several types of machine tools, including lathes, grinders, and milling machines; (2) paper-machine drives; (3) rubber-calender drives; (4) conveyor drives; (5) textile-range drives; (6) steel-mill auxiliary drives; (7) testing-equipment drives.

In many applications the entire motor power is supplied through electronic rectifiers of the controlled type. The thyratrons are then usually controlled from vacuum-tube control circuits. Most applications involving electronic power supplies have been made in the low- and medium-horsepower field, between $1/4$ and 30 hp. Although some applications of electronic power supplies have been made as high as 300 hp, the more conventional power converter in the large size is a motor-generator set. In some motor-generator-set applications, thyratron rectifiers are used to control the field of the generator or the motor or both, as necessary. High-vacuum tubes are then generally used to control the thyratrons. In other applications electronic control is limited to the use of vacuum tubes for controlling the low-power fields of pilot exciters, such as the General Electric amplidyne, which in turn control the generator or motor-field power.

Figure 14 is a typical electronic control circuit which provides adjustment of the armature voltage of a d-c motor by thyratron phase-shift control to obtain speed control. Figure 15 shows pictorially the essential components of a complete adjustable-speed drive. This drive provides the following features: (1) adjustable motor speed over a wide range from an a-c power supply; (2) accurate speed regulation; (3) current-limit acceleration to a preset speed; (4) compensation for the internal motor *IR* drop, regardless of motor load; (5) normal motor torque at the instant of closing the motor-armature contactor.

The motor field is supplied by a non-controlled rectifier (not shown). D-c control power is obtained from a small control-power rectifier (not shown). The a-c winding of the saturable reactor *SR*, the d-c winding of which is shown, is connected in the thyratron phase-shifting bridge to adjust the rectifier output voltage in proportion to the amount of current in the d-c winding. Tubes 1 and 2 are voltage-regulating glow tubes. The speed-control potentiometer *1P* which is connected across the lower glow tube (tube 2) provides a reference voltage against which the motor-armature voltage is compared.

Conventional glow-tube voltages are 75, 105, and 150 volts. The d-c motor voltages are commonly 230 volts. Therefore a voltage divider is connected between lines 24 and 26 so that the voltages between lines 22 and 7 will be about equal to the reference voltage when the motor-armature voltage is 230. The voltage from lines 22 to 7 includes an *IR* compensation voltage which will be discussed later. For the moment, however, assume that the *IR* compensation potentiometer, *3P*, is turned counterclockwise so that 26 and 7 are at the same potential. Any difference in voltage which exists between the slider of the speed control and point 22 is applied to the grid of tube 4. A capacitor is connected from grid to cathode of this tube to filter the rectifier output voltage before it is applied to the grid.

If the armature voltage is proportionally lower than the speed-control slider voltage the grid of tube 4 will be negative with respect to the cathode and therefore tube 4 will conduct no plate current. This results in a lower voltage drop through resistor $2R$ and, consequently, a less negative voltage on the grid of tube 3, which increases the plate current of tube 3.

The increased plate current which flows through the d-c winding of *SR* increases the rectifier output voltage until the grid voltage of tube 4 differs from the cathode only by the voltage required to produce the desired rectifier output voltage. A small grid-voltage change will result in a range of output voltage from 0 to maximum. If the speed-control slider is moved, therefore, there exists momentarily a large differential voltage applied to the grid. This will turn the power tubes entirely on or off, as the case may be, until

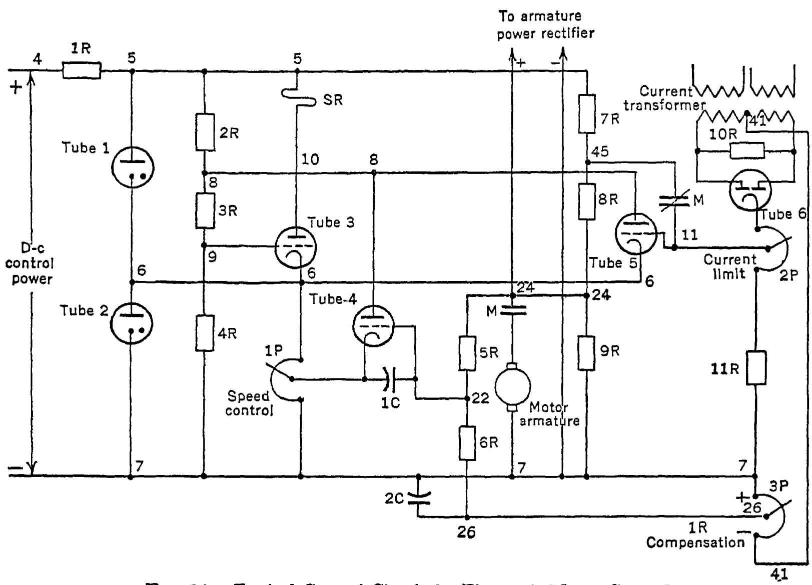


FIG. 14. Typical Control Circuit for Electronic Motor Control

the armature voltage has reached the new level. Once it has reached the new level only a small difference between reference and signal voltage is needed to control the rectifier.

If the field current of a d-c motor is held constant the no-load speed is proportional to the applied voltage. When mechanical load is applied to the motor shaft, however, the armature current is increased. This results in a voltage drop in the internal motor-armature winding which reduces the speed in proportion to the voltage drop. A small motor may have a drop of 23 volts at full load. This is 10 per cent of the rated armature voltage of a 230-volt motor. At rated applied voltage then there will be a speed regulation of 10 per cent between no load and full load. At 10 per cent of speed and no load, however, the applied voltage is only 23 volts. If this voltage is held constant the speed will be reduced to 5 per cent at 50 per cent of load and further reduced to 0 speed at 100 per cent load. This is obviously not a satisfactory method of operating, particularly at low speeds.

Compensation for the *IR* drop is obtained in the circuit Fig. 14 in the following manner.

One primary of a current transformer which has two electrically separate primary windings is connected in each of the two thyratron rectifier-tube anode circuits. The secondary of this current transformer is connected to the rectifier tube 6. A d-c voltage which is proportional to the armature load current therefore appears across potentiometer $2P$, resistance $11R$, and potentiometer $3P$. The polarity of the voltage across potentiometer $3P$ is shown. Assume now that the motor is running at no load at any desired speed and that the *IR*-compensation potentiometer $3P$ is turned clockwise. Since no armature current is flowing there is no voltage across potentiometer $3P$ and therefore the voltage from 22 to 7 is proportional only to the armature voltage. Now, if motor load is applied, line 26 will become more negative than line 7. As a result the grid of tube 4 will also become more negative. This change in grid voltage acts through the amplifier to apply more armature voltage. The armature voltage is thus raised to compensate for the resistance drop in the motor due to load. Potentiometer 3*P* can be adjusted to compensate for the resistance drop in different motors and therefore to hold the speed constant with load changes. Capacitor 2*C* is used to filter the rectifier output voltage.

The current transformer and rectifier circuit also provide a voltage signal to prevent overcurrent which would result in exceeding the rated peak current of the power tubes, the commutating limit of the motor, and perhaps the torque limit of the driven load.

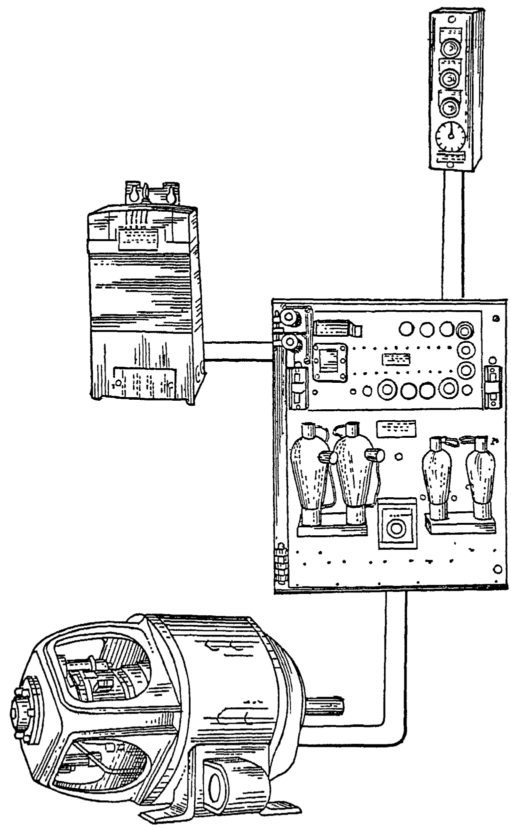


FIG. 15. Pictorial Elements of Electronically Controlled D-c Motor Drive (Courtesy General Electric Co.)

Assume that the speed control is set at a high-speed point and the motor-armature contactor *M* is closed. Tubes 4 and 3 will now control the thyratron rectifiers to give full voltage output. This voltage would result in approximately 10 times normal armature current at the first instant. The voltage between points 7 and 11 is proportional to the armature current. If the armature current is below a normal value (approximately 150 per cent of full-load current), the grid of tube 5 will be negative with respect to the cathode. If, however, it exceeds a normal value, the grid of tube 5 will become less negative and the plate current in tube 5 will control the output voltage of the rectifier (in the same manner as tube 4 does) to hold the accelerating current at a predetermined maximum value.

When the armature contactor *M* is open, the rectifier output voltage is still controlled by the speed-control potentiometer. If the speed-control slider were set for a high-speed position there would still be an initial high current impulse when the armature contactor first closed. To prevent this condition a voltage divider consisting of resistors 7*R*, 8*R*, and 9*R* is connected across the regulated d-c voltage. Also the rectifier output voltage is connected to the divider at point 24. When the armature circuit is open, the normally closed interlock on *M* connects points 45 to 11. Now, as the rectifier voltage is increased, tube 5 draws plate current which holds the rectifier output voltage at a level which will result in approximately normal armature current and therefore normal motor torque at the instant the armature contactor is closed.

Many other features can be obtained by circuit variations. Some common variations are: (1) control of the motor-field voltage by thyratron rectifiers; (2) control of motor- or generator-field current; (3) more accurate speed control by means of a small permanent-

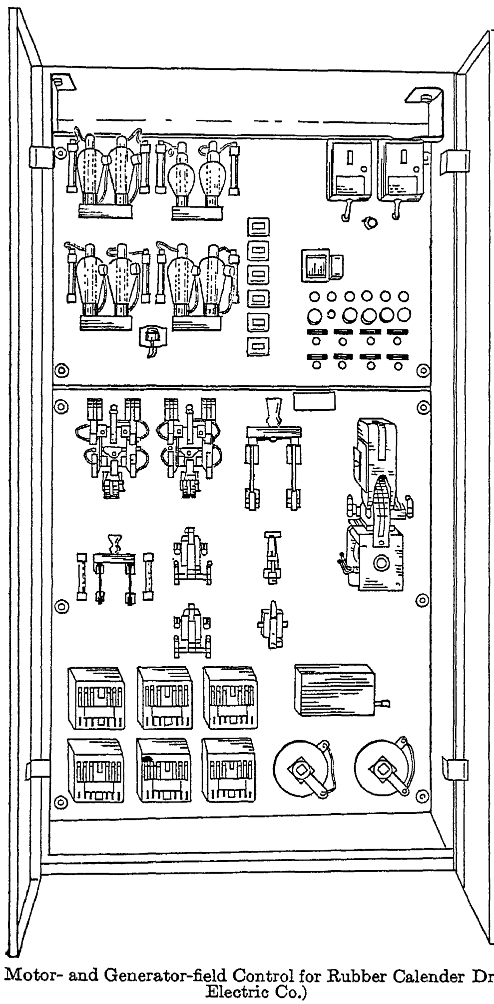


FIG. 16. Electronic Motor- and Generator-field Control for Rubber Calender Drive (Courtesy General Electric Co.)

magnet-type d-c generator mounted on the motor shaft and used as a speed signal; (4) reversing by current-limit regenerative means; (5) speed programming; (6) timed acceleration in lieu of current-limit acceleration; (7) motor-speed control in accordance with external signals.

Figure 16 shows a control panel for a rubber-mill calender drive which includes electronic control of the fields of a generator and a 125-hp motor.

14. SIDE-REGISTER POSITIONING CONTROL

Servomechanism is a term often applied to the general field of follow-up or positioning control. Side-register positioning control is one of the many examples of positioning control which could be selected from this broad, general field. Side-register control is applied to steel-mill "coilers," to papermill "reels," and to textile-mill "beams" to obtain a smooth surface at the edge of the finished-material roll. It is also used on slitters for insulating tape, linoleum, and the like to guide material going from a reel into the slitting knives so that only a certain amount of undesired edge will be trimmed, with a minimum of waste beyond the slitter.

One frequent use is an electronic side-register positioning control applied to a paper slitter. In this application it is desirable to hold the edge of the material at a given point with respect to the slitting knives. A photoelectric scanner, including a light source, a phototube, and a preamplifier, is focused on the edge of the paper. The photoelectric preamplifier supplies a signal to the electronic control panel, which in turn operates an alignment motor to move a large roll of paper, as well as the nip roll, back or forward in order to obtain a fixed position of the edge of the paper. Different types of side-register positioning-control equipment are available to suit varying requirements, such as variation in contrast between the material and its background, range of colors to be used, alignment-motor horsepower, accuracy of positioning, and maximum correction rate required.

The circuit diagram shown in Fig. 17 is that of a simple on-off type control. This control with its associated reversible a-c motor and mechanical brake will meet the requirements of applications where: (1) there is a reasonably large light contrast between the

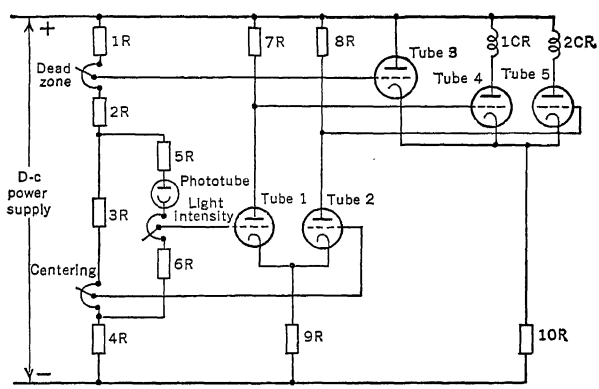


FIG. 17. On-off Type Photoelectric Side-register Positioning-control Circuit

material to be positioned and its background or where direct light transmission can be used; (2) the required accuracy is not greater than plus or minus $^1/_{16}$ in.; (3) the required rate of correction is not greater than 5 in. per min. The correction motor in this case is a reversible a-c motor which is operated in the forward or reverse direction by magnetic contactors, which are in turn operated by relays *1CR* and *2CR*. If there is no light shining on the phototube, the grid of tube 1 will be negative. If tube 1 is not drawing current, the cathode will go slightly negative, but, in doing so, since the cathode of tube 2 is connected to the cathode of tube 1, tube 2 will be turned on. The lack of current in tube 1 will cause tube 4 to conduct and energize relay *1CR*. Since tube 2 is conducting, tube 5 will be non-conducting and relay *2CR* will be de-energized. If a large amount of light is now applied on the phototube, the grid of tube 1 will be positive and tube 2 will consequently be turned off. This will result in the opposite operation of relays *1CR* and *2CR*.

The light-intensity adjustment can be turned clockwise to decrease the intensity that will cause the relays to operate. For a given light intensity, the centering adjustment will make the transition in the relays occur when approximately half of the light beam is intercepted or reflected, depending on the method of light transmission. It is obvious that, if there were no dead zone between the operation of the two control relays, the equipment would be continually hunting from one side to the other. The dead-zone adjustment provides an adjustable amount of distance over which the material can travel without actuating the control relays.

In applications where continuous operation of the correcting motor is desirable, the system just described is not suitable because of the maintenance of the mechanical parts. In these applications the photoelectric amplifier can control two half-wave thyratron rectifiers which are connected in inverse parallel to operate a shunt-wound d-c motor in the forward or reverse direction. For correcting motors of $^3/_4$ hp or larger, vacuum tubes are generally used to operate the field of a pilot generator which will in turn supply power to operate the correcting motor in the forward or reverse direction. Where wide ranges of light intensities and small light differentials are to be encountered, a more elaborate control equipment involving a rotary-lens-type scanning head is often used.

15. PROCESS CONTROLS

Electronic control has recently made its way into the field of process instrumentation and control, where it is now frequently used in measuring, recording, and controlling such process variables as temperature, pressure, flow, pH, moisture, and the like. An electronic control unit included as a part of a complete potentiometer controller is shown semischematically in Fig. 18. The battery supplies a standard reference voltage to the slide

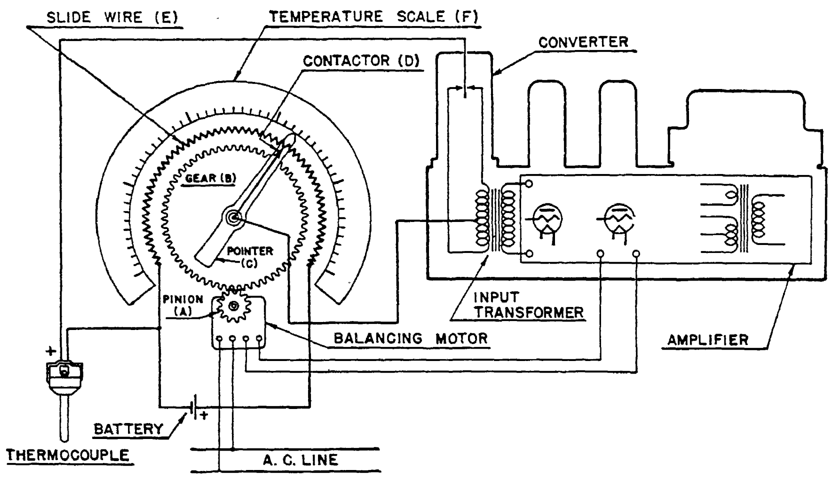


FIG. 18. Semischematic Diagram of Continuous-balance Potentiometer Controller (Courtesy Brown Instrument Co.)

wire. The voltage produced by the thermocouple is compared with the voltage at the slider. Since the voltage magnitudes are so small, however, it is not suitable to apply this voltage difference directly to a d-c amplifier. Instead, it is connected to a vibrating-type inverter which is synchronized with the power-supply frequency to produce a-c voltage impulses of a given phase relation in the amplifier tubes. If a difference exists between the thermocouple voltage and the voltage of the standard, this signal is amplified by an a-c voltage amplifier and further by means of vacuum-type power-amplifier tubes to control the power applied to a small reversible-type a-c balancing motor. The reversible balancing motor repositions the slider until the thermocouple voltage matches that of the battery. Thus, the pointer is positioned to show the temperature which exists at the thermocouple location.

If pressure is to be measured or controlled, the pressure-sensitive element can be a variable reactance. This can be used in an a-c bridge circuit, and the mechanical inverter is not necessary.

16. SYSTEM STABILIZATION

In any electromechanical system there is a time lag between the application of a correction signal and the final corrected value; for example: (1) if a given voltage is applied to the field of a generator, the inductance of the field prevents the field current and therefore the generator output voltage from building up instantaneously; (2) if a given voltage is applied to the armature of a d-c motor the mechanical inertia prevents the motor speed from instantaneously reaching its new value; (3) if a correction voltage is applied to the correcting motor of a servomechanism positioning control not only must the correcting motor accelerate its own inertia plus the inertia of the connected load but it must also run at rated speed until the new position is reached—then it must decelerate.

Assume that a sensitive and instantaneously operating electronic voltage regulator is applied to a d-c generator. If the generator output voltage is lower than that called for by the voltage-adjusting potentiometer, the regulator will immediately apply the maximum field voltage to attempt to correct the output voltage. The output voltage, however, will increase slowly because of the field inductance. However, when the output voltage reaches the desired value the electronic regulator immediately applies the correct field voltage to hold the desired output voltage. Such a system may be slow in response, but it is stable in its operation.

If on the other hand it is assumed that the electronic regulator has internally an inherent time lag, the operation of the entire system will be quite different. If the generator voltage is low, the regulator will again apply maximum field voltage. When the generator voltage reaches the desired point, the regulator will not immediately apply the correct field voltage because of the inherent time lag in the regulator. The correct field voltage will not be established until the output voltage has exceeded the desired value. Now, however, the regulator receives a signal which will result in a weaker field than normal. As a result, the system may be unstable and continue to oscillate above and below the desired voltage level.

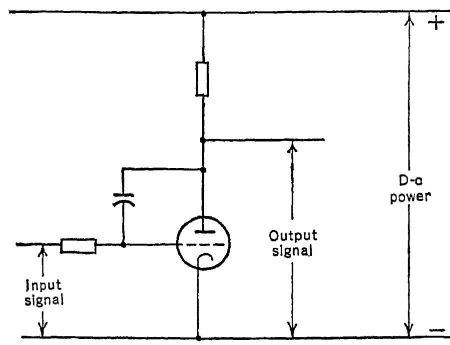


Fig. 19. Method of Adding a Predominant Time Lag to Reduce System Hunting

The frequency of oscillation or hunting is determined by the time constants of a system. If long time constants are involved, the hunting will be at a low frequency; if short time constants are involved, the hunting will be at a high frequency. In order for hunting to exist in any system, two conditions must exist: (1) the total time lags of the entire regulating system must result in the regulating-voltage signal being fed back to the regulator input 180° out of phase with the normal sense of the correction voltage; (2) at the frequency at which the total time lags add up to 180°, the overall system amplification must be equal to or greater than 1.

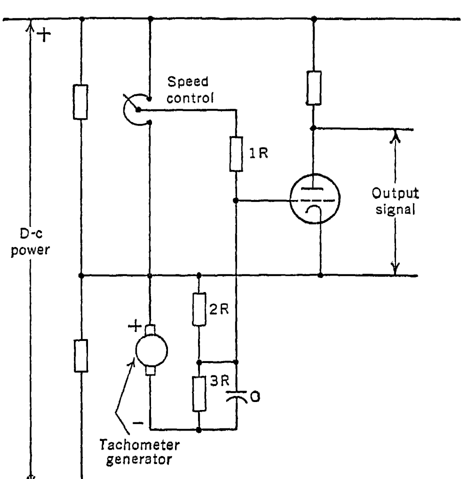


Fig. 20. Method of Adding a Derivative or Rate-of-change Signal to Reduce System Hunting

Obviously hunting is undesirable in most systems. Therefore means must be established to eliminate it. Several methods are in common use.

As was pointed out earlier, a system with a single time constant will not hunt. (Although this is true in general there are single-lag systems which will hunt.) Hunting in a system having two or more independent time lags can be reduced by artificially adding a time constant which is sufficiently greater than the others that it predominates. Such a system will approach a single-lag system in performance. Figure 19 shows how a lag can be added to a d-c amplifier circuit. The grid-to-plate capacitor charges or discharges through the series grid resistor at a rate dependent on the circuit constants. The action of the grid therefore lags behind the input signal.

Since the overall amplification of the system must be at least 1 at the hunting frequency, a simple way of reducing hunting is to reduce the system amplification. Conversely, the higher the system amplification is, the more likely the system is to hunt.

Often rate of change or derivative signals may be added to the input signal to reduce hunting. Figure 20 shows a typical circuit used in a speed regulator. The steady-state grid-voltage signal from the tachometer-generator which is connected to the motor shaft is determined by resistors $2R$ and $3R$. During a change in speed, however, the capacitor C supplies a grid voltage which is proportional to the rate of change of speed and thus provides a grid signal which is ahead in time relation to the speed signal.

Rigorous solutions to stability problems are often complex. Furthermore, assumptions of system constants or simplifying assumptions often are proved to be in error when equipment is installed. As a result most systems are stabilized by empirical methods.

BIBLIOGRAPHY

Chin, P. T., and E. E. Moyer, Principles of Grid Control for Thyratrons, *A.I.E.E. Technical Paper* 45–62 (January 1945).

Chin, P. T., and E. E. Moyer, A Graphical Analysis of the Voltage and Current Wave Forms of Controlled Rectifier Circuits, *Electrical Engineering*, July 1944.

Chute, G. M., *Electronics in Industry*, McGraw-Hill (1946).

Chute, G. M., Fundamentals of Industrial Electronics, *Steel*, April 3–May 22, 1944.

Chute, G. M., *Resistance Welding Control*, McGraw-Hill.

Cockrell, W. D., Grid Control of Gas Filled Tubes, *Electronics*, June 1944.

Cockrell, W. D., *Industrial Electronic Control*, McGraw-Hill (1944).

Cooper, B., Better Welds through Regulated Welding Current, *Welding Journal*, January 1944.

Dalton, B. J., Electronic Motor Control, *Proceedings of National Electronics Conference*, 1944.

Dalton, B. J., Electronic Motor Control, *Gen. Elec. Rev.*, May 1945.

DeBlieux, E. V., Characteristics, Design and Application of Rectifier Transformers, *Gen. Elec. Rev.*, October, November, December 1937.

Hall, James H., Transformer Calculations for Selenium Rectifier Applications, *Electrical Manufacturing*, February 1946.

Hazen, H. L., Theory of Servo-Mechanisms, *Journal of the Franklin Institute*, September 1943.

Kloeffler, R. G., *Principles of Electronics*, Wiley (1942).

Leigh, H. H., and H. L. Palmer, Inverter Action on Reversing of Thyratron Motor Control, *Electrical Engineering*, April 1944.

M. I. T. Staff, *Applied Electronics*, McGraw-Hill.

Moyer, E. E., Electronic Control of D-c Motors, *Electronics*, May, June, July, September, October 1943.

Palmer, H. L., M. E. Bivens, and S. A. Clark, Electronic Welding Control, *Electronics*, August, September, October 1943.

Reich, H. J., *Theory and Applications of Electron Tubes*, McGraw-Hill (1939).

Smith, E. S., *Automatic Control Engineering*, McGraw-Hill (1944).

SECTION 22

AIDS TO NAVIGATION

RADIO AIDS TO AIR NAVIGATION

By Henry I. Metz

RADIO AIDS TO MARINE NAVIGATION

By M. K. Goldstein

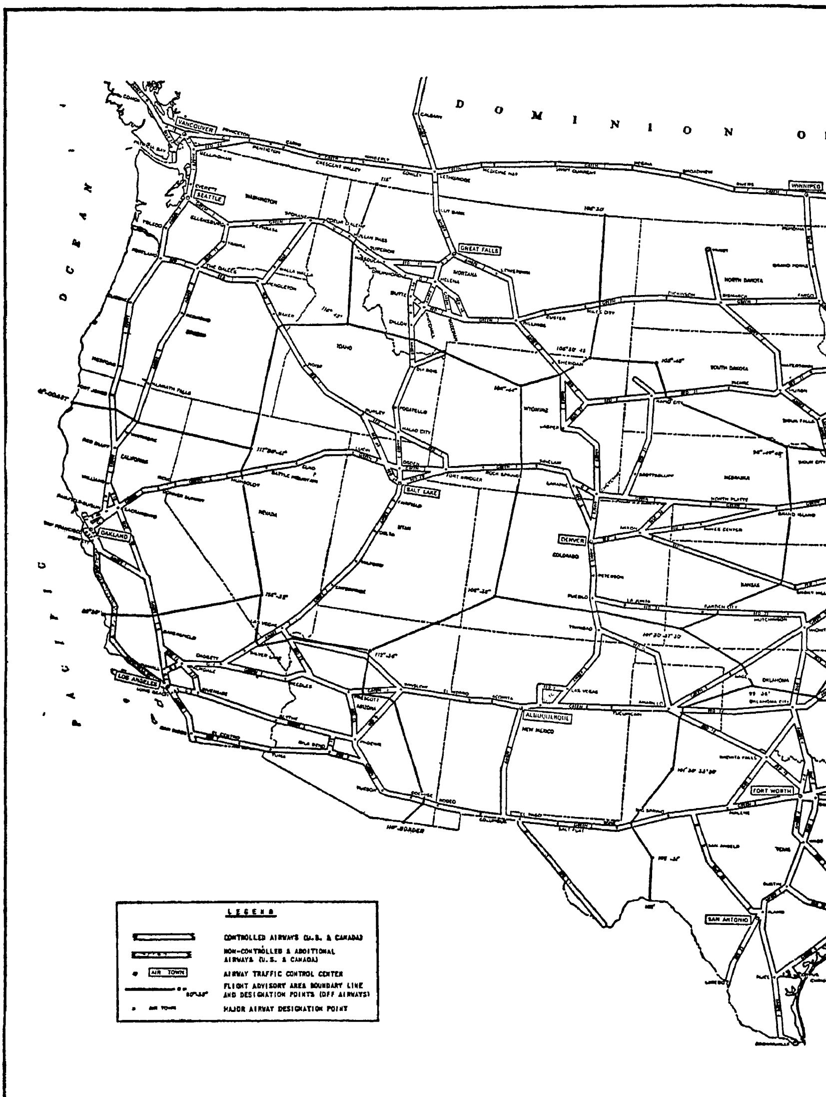
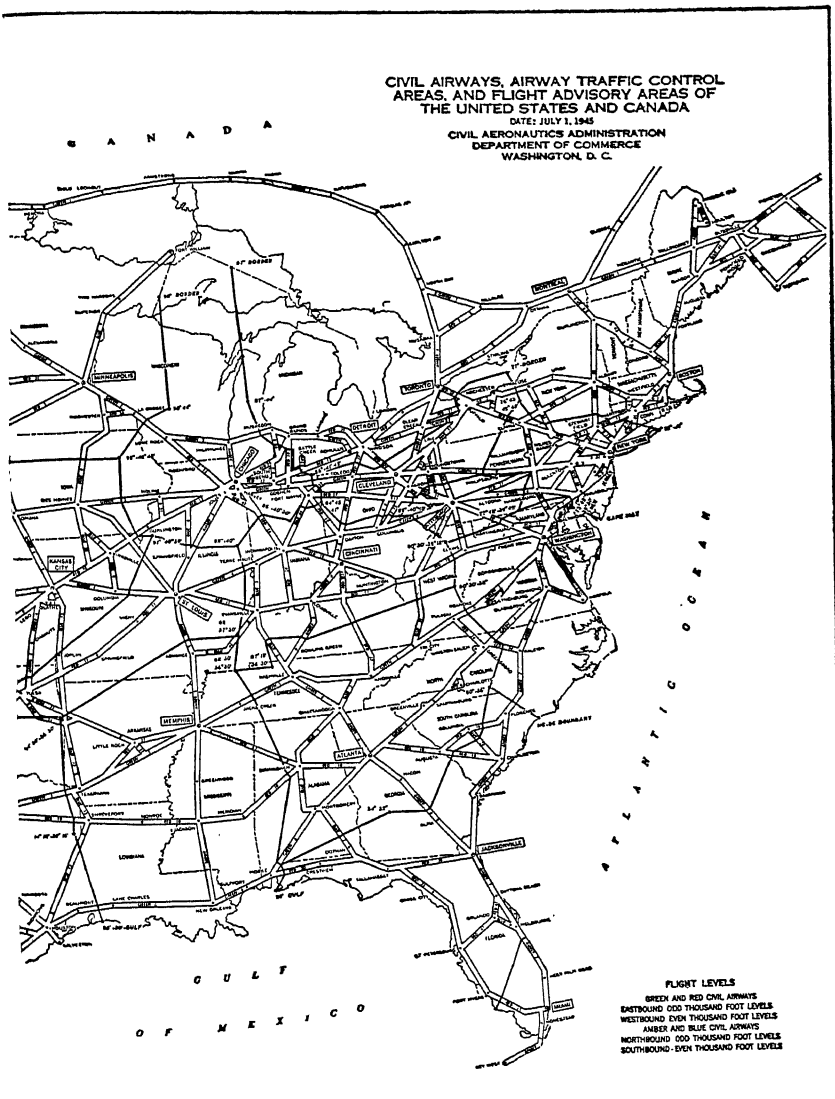


FIG. 1. The Present 35,000 Miles of Federal and Civil Airways. The airway



traffic control boundaries and centers are also shown. (Courtesy CAA.)

AIDS TO NAVIGATION

RADIO AIDS TO AIR NAVIGATION

By Henry I. Metz

1. INTRODUCTION

Thirty-five thousand miles of federal airways exist today in the airspace over the United States. This mileage, shown in Fig. 1, is constantly and very rapidly increasing. Information about the condition of the airway, weather ahead, and other traffic is available constantly to the pilot by means of automatic radio aids to navigation and two-way radio voice communications.

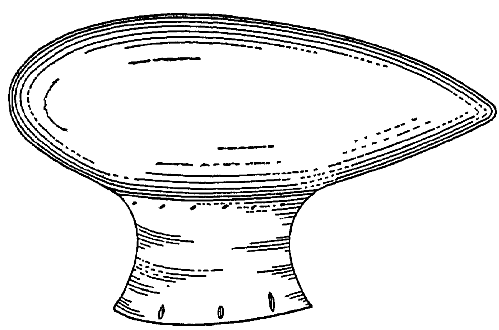


Fig. 2(*a*). View of Aircraft Automatic Direction Finder Equipment Showing the Loop Antenna (Courtesy Bendix Aviation Corporation)

The establishment, operation, and maintenance of the airways are among the functions of the Civil Aeronautics Administration (CAA) under the U. S. Department of Commerce.

When the federal airways program was started in 1926, the development of a reliable radio communication and guidance system was undertaken. Basically, tracks were established in the airspace by overlapping keyed radio patterns, the points of overlap being interpreted aurally by the pilot, who wears a pair of headphones. Although originally produced by a pair of crossed loop antennas, the tracks are today made by an Adcock array of vertical radiators which give greater night-time stability. The transmitting station is called a "radio range station." It produces four tracks called "courses," all of which emanate radially in predetermined fixed directions. There are 399 range stations now in operation. The coordinated alignment of a series of range courses constitutes an airway.

A system of markers has been added to the range courses. These are vertically directed signals at 75 Mc, received on a special receiver in the airplane and connected to a signal light on the instrument panel. The light operates only when the airplane is over the marker station.

An automatic direction finder (ADF), as shown in Fig. 2, is carried today by most commercial airplanes and is required (in some form) by CAA on scheduled airliners. Its pointer indicates the direction to the station tuned in. It may be used as an aid in flying the range course or to determine position by taking bearings on two or more stations.

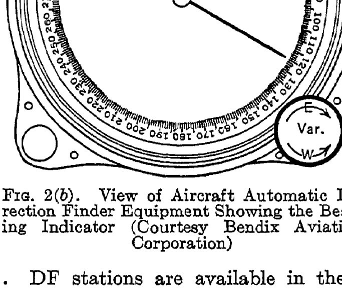


Fig. 2(*b*). View of Aircraft Automatic Direction Finder Equipment Showing the Bearing Indicator (Courtesy Bendix Aviation Corporation)

Countries other than United States, Canada, and Australia have based their traffic operations on ground station direction finding (DF). DF stations are available in these countries to give bearings to the aircraft calling, just as is now done in locating position on ships at sea.

Airborne direction finders have also been used extensively in other countries in cooperation with high-powered, non-directional, transmitting stations on the ground. A system called "radio-phare" has been employed in which three spaced ground stations transmit in time sequence on the same frequency. Without readjustment of the airborne ADF receiver, three bearings can be quickly obtained from the radio-phare. The German 33-Mc "Lorenz" instrument landing system (localizer, glide path, and markers), in particular the localizer, was in quite general use before the war. Its localizer, using interlocked aural dots and dashes (similar to the U. S. Aural AN Radio Range) to differentiate left from right in approaching the runway, was used for distances up to 100 miles in cases where the fixed course alignment agreed with the desired flight direction.

The CAA is engaged in a program of converting all the federal airways aids to static-free VHF, except that a few strategically located low-frequency range stations will be retained (with increased power) to serve for long cross-country flights. In addition to converting to VHF, all future ranges will be of the "visual" type (i.e., using pointer instruments instead of headphones), and will have omnidirectional courses (i.e., instead of only four range courses, the pilot may select any radial track, or course, from the station) to take care of increased numbers of airplanes and airports. Instrument landing systems, developed previously and now installed at 110 airports, will be used at once to relieve airport traffic congestion in bad weather. Radar will be used soon as a traffic surveillance device and, perhaps later, as a direct means of controlling traffic. Pulse techniques will probably be applied immediately in new meter-type distance-measuring equipment (DME) and later in anticollision devices. Long-distance flights across water will be guided by relatively low-frequency aids such as omnidirectional, Loran, Sonne, or other systems.

Automatic flight, controlled by radio signals transmitted from the ground, has already been demonstrated to be more accurate than human-pilot-controlled flight. It will undoubtedly be used by all commercial aircraft in conjunction with automatic computers, to permit controlled, track-type flying *in any direction* regardless of station position.

2. TERMINOLOGY AND DEFINITIONS

Certain terms associated with radio aids to air navigation in engineering discussions are listed below with their definitions:

Radio Range. Any CW radio station whose radiation inherently produces directional courses, or tracks, fixed in their relation to the earth's surface and independent of aircraft heading. A radar range is similar but uses pulse radiation.

Marker. Any station having limited or directed radiation used to give an aircraft its position along a range course.

Heading. Direction (azimuthal and clockwise from north) in which aircraft is pointing. It agrees with direction of flight if there is no crosswind.

Bearing. Azimuthal angle (from north, clockwise) of line between fixed ground station and airplane, or vice versa. It is generally necessary to state where the bearing is from and to, in order to avoid ambiguity.

Track. Actual direction of motion of aircraft with respect to earth's surface, expressed in degrees of azimuth (from north).

Beacon, Non-directional. A radio transmitting station whose radiation is essentially uniform in all directions or which does not use directional radiation characteristics to convey intelligence.

Beacon, Omnidirectional. A beacon whose directional or other radiation characteristics cause it to give information equally in all directions.

Course Sharpness. The relation between angular displacement from course and deflection of the pointer of the indicating instrument, usually expressed in angular degrees of displacement required to give full-scale-left to full-scale-right pointer movement. This sharpness is generally a function of ground-station pattern sharpness and receiver gain setting. Localizers are generally used with 4° to 5° sharpness.

Pattern Sharpness. The difference in pattern amplitudes at a given angular displacement from the equal or on-course line, usually expressed in decibels per 1.5°. Standard figure-of-eight patterns (LF aural ranges) give 0.45 db per 1.5°. The CAA localizer patterns give about 5 db sharpness.

Clearance. The db difference in patterns producing a course, at angles other than those containing the on-course. High clearance is desirable so that the indicator will remain fully deflected everywhere except at the course.

Multiples. Extra or abnormal courses resulting from zero or negative pattern clearance or from severe reflection of signal from buildings, trees, etc.

Bends. Angular deviation or distortion of the on-course signal from a true, straight radial line from the station. Generally, bends are produced by reflection of the signal from buildings or wires near the transmitting station. Bend magnitude is proportional to the ratio of reflected signal to direct signal amplitude and inversely proportional to pattern sharpness. Apparent bends can be caused by poor receiver AVC action, allowing inferior circuit components to unbalance as signal strength changes with distance.

Scalloping. The irregular or wavy shape of an antenna field pattern caused by reflection of the signal from ground objects. Scalloping is evidenced by the periodic hesitation in the course indicator movement as the airplane is flown across the course. Scalloping indicates bends when it occurs near the on-course line. When scalloping is severe, multiple courses are produced.

Wiggles. Rapid, random, and erratic movement of the course-indicating pointer, generally caused by combined signal from several reflecting objects, especially trees. Poor electrical connections and noise also cause wiggles. Wiggles generally do not alter the average course line and can therefore be filtered out of the indicator.

Pushing (or **pulling**). Displacement of indicated course with heading of airplane. Term is derived from observation in cross-course flights that the course was apparently "pushed" ahead of the airplane. Pushing is caused by the radiation of impure polarization (vertical in a horizontally polarized system, and vice versa). *Attitude effect,* in which the indicator shifts with airplane roll, or pitch, is similar in cause to pushing.

Distance Range. Distance in miles from the station at which useful signal is lost, or where course sharpness decreases (loss of AVC). For VHF and above, the line of sight range generally prevails.

3. RADIO AIDS IN THE FEDERAL AIRWAYS SYSTEM TODAY

The present system consists of the following facilities or services:

A. Radio ranges, LF (four-course aural with simultaneous voice).
B. Radio ranges (visual two-course, VHF).
C. Radio markers (75 Mc).
D. Automatic direction finder (ADF) receivers.
E. Communications (HF and VHF).
F. Air traffic control and weather reporting.

Each of these is explained separately below.

RADIO RANGES, LF (four-course aural with simultaneous voice). The radio range of today is a 400-watt, highly developed, four-course facility on which practically all civil air navigation is based. It is not a perfect device but is simple and effective for distance ranges up to several hundred miles. Its irregularities are so well known that many of them appear to the trained pilot as an asset. The change in conductivity of the earth along the course, in some places, usually in mountainous terrain, causes bends and multiple courses to appear. The location of these multiples is known and is plotted on charts. Night effect, which is variable, has been minimized by the replacement of loop-type transmitting stations with Adcock vertical tower systems. Two highly important objections to the range are (1) interference by atmospheric and precipitation static, and (2) interference from other stations in the allotted frequency spectrum.

The present four-course Adcock antenna system consists of five steel towers about 130 ft high. Four towers are placed on the corners of a square; diagonally opposite towers constitute a pair and are about 600 ft apart. The fifth tower is at the center. The pairs are connected, but with reversed phase, so that they, respectively, radiate figure-of-eight horizontal patterns. The two pairs are connected to a crystal-controlled transmitter through the equivalent of a single-pole, double-throw relay called a "link circuit relay." The relay is operated by a motor-driven keyer unit so that one pair of towers gives a series of dot-dash (*A* in code) characters. The back contact of the relay causes the reciprocal character, dash-dot (*N* in code) to be keyed into the other pair of towers. The schematic diagram is shown in Fig. 3.

The two figure-of-eight patterns thus contain reciprocal or interlocking characters. To an observer with a receiver in space the character heard would depend on position around the station. In a position where the patterns overlap, that is, where the *A* and *N* patterns are of equal amplitude, no character would be observable because they are interlocked and the signal is a steady tone. The patterns and courses are shown in Fig. 4.

Actually, the radiation from the two pairs of towers is unmodulated carrier energy and therefore inaudible. The center (fifth) tower is connected to a second crystal-controlled transmitter, but its carrier frequency is 1020 cycles below that used for the diagonal towers. Consequently, in combination, the receiver produces a 1020 cycle-output resulting from

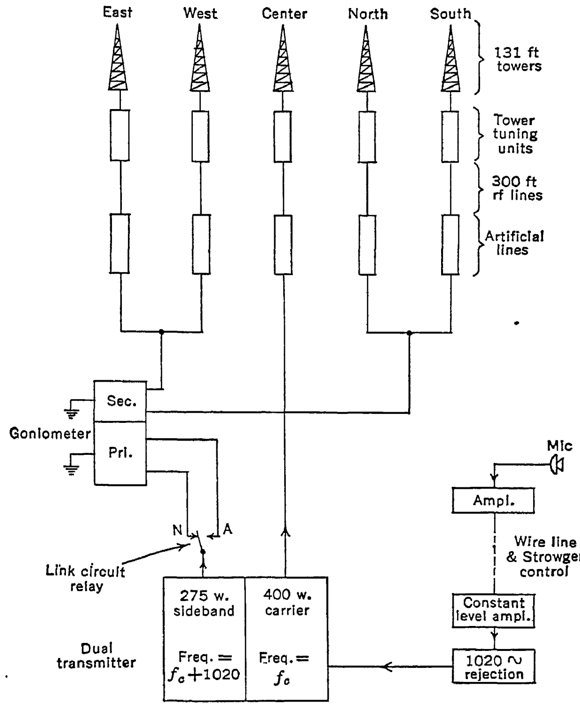


FIG. 3. Schematic Diagram of Four-course Aural LF Radio Range Showing Simultaneous Voice

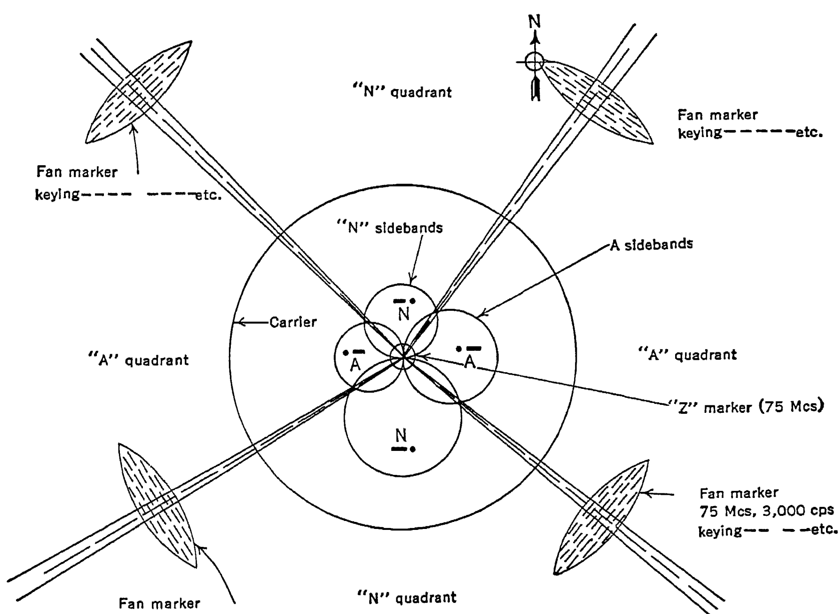


FIG. 4. Patterns and Courses of LF Aural Radio Range Showing Flexibility of Course Displacement and Position of 75-Mc Markers

the beat between the two carriers. In effect, the center tower radiates the carrier and the receiver is tuned to it. The signal from the side towers constitutes sideband energy. The power of the side tower transmitter is adjusted to give 30 per cent of that of the center tower carrier signal along the course. The remaining 70 per cent carrier is used for transmission of voice modulation. The airborne receiver generally contains a combination band-pass, band-rejection filter so that the pilot may select voice or range signals without interaction.

The course of the aural *AN* range is normally about 3° wide. The width depends upon headphone level and the pilot's hearing ability. The change from on-course to full off-course is gradual; that is, a "twilight" zone exists to either side of the course which permits the pilot to estimate (very approximately) his nearness to the course.

Normally, four courses are produced at right angles to each other. All four courses may be rotated equally, clockwise or reverse, by turning a goniometer through which the outside antennas are connected. Or, if other than right-angle courses are required, as shown in Fig. 4, the relative phase of the current in opposite towers is varied by means of artificial lines. The total line length between towers is maintained at an optimum value for maximum phase and current stability. Under optimum line conditions, the detuning of rain, ice, and snow on one antenna reacts to produce an equal effect on the other antenna, giving greatest stability of courses.

RADIO RANGES (visual two-course, VHF). The development of a two-course VHF radio range was started in 1928 to overcome the static, congestion, and dangerous quadrant ambiguity problems existing on the conventional four-course, low-frequency, aural range. Now several airways are operating (temporarily on the 110-Mc frequency band) with this type of facility, pending conversion of the entire federal airways to the VHF omnidirectional system.

The complete designation of the two-course range is: "VHF two-course visual radio range with quadrant identification." This designation signifies particularly that the quadrant ambiguity of the four-course type has been eliminated. Actually, the quadrant identification comes through the superposing of two aural courses on the same visual range, but with these aural courses at right angles to the visual. This facility, therefore, consists of two visual and two aural courses, as shown in Fig. 5. Normally, the visual course is

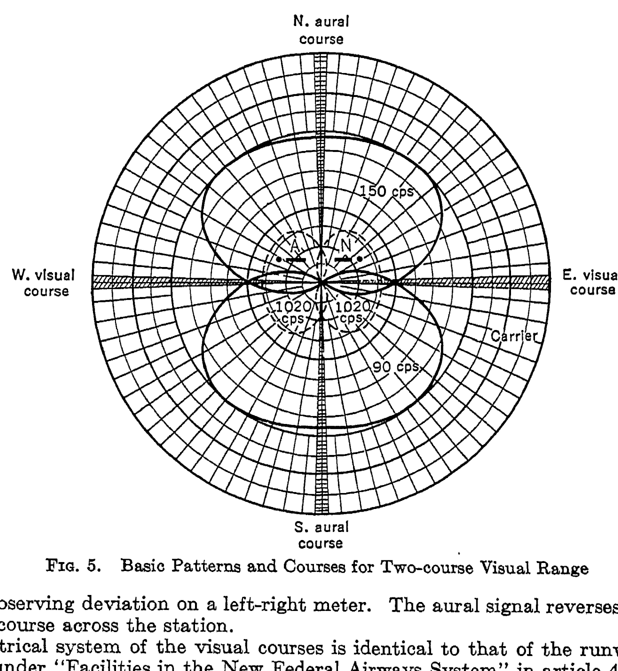


FIG. 5. Basic Patterns and Courses for Two-course Visual Range

flown by observing deviation on a left-right meter. The aural signal reverses when flying the visual course across the station.

The electrical system of the visual courses is identical to that of the runway localizer described under "Facilities in the New Federal Airways System" in article 4. The aural courses are laid down as described above in the discussion of aural radio ranges. Requiring less course sharpness than the localizer, its antenna array consists simply of three horizontal loops for the visual courses and two additional loops for the aural. The central loop serves for both the aural and visual courses (see Fig. 6).

A simultaneous voice feature is incorporated in this range facility by modulating the carrier fed to the center loop with voice. The total modulation capacity of the transmitter is divided approximately as follows: 90 cps 20 per cent, 150 cps 20 per cent, 1020 cps 10 per cent, voice 40 to 50 per cent.

The sharpness of the visual course is dependent upon the transmitted pattern shape and the signal voltages delivered to the receiver indicating circuit. In practice, the

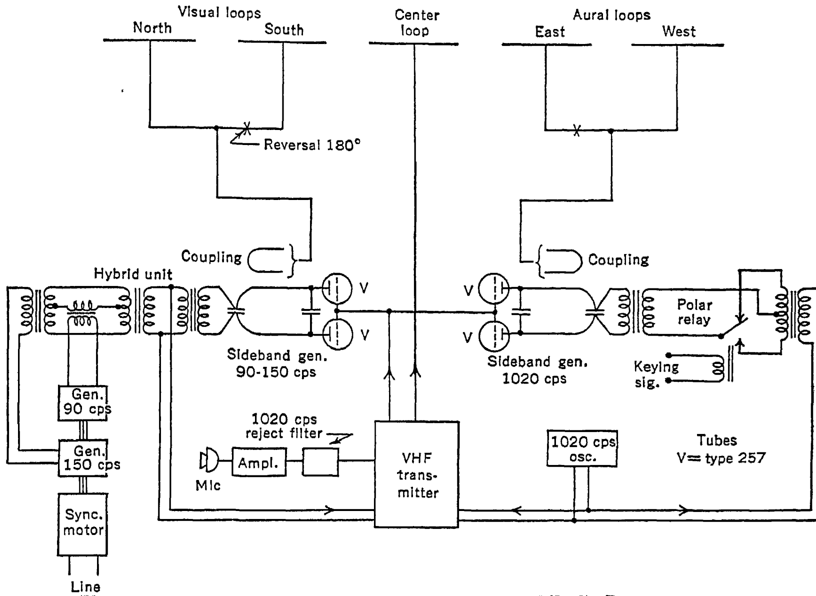


FIG. 6. Diagram of VHF Two-course Visual Radio Range

receiver gain is adjusted to give full scale left to right deflection of the indicator for a 20° azimuthal displacement of airplane (plus or minus 10° from the course). The aural course sharpness depends upon the patterns and also the pilot or observer. On this range it is about 2°.

The pilot selects the voice broadcasts or 1020-cps aural signals of this range through the standard range-voice filter described for the four-course range.

This facility is subject to reflection and propagation phenomena characteristic for 110 Mc VHF. Objects near the station, such as wires, buildings, and trees, reflect the signal and give multiple path transmission to the receiver, causing scalloping of the patterns and sometimes bends or multiple courses. In a moving airplane the random reflections of trees causes wiggles in the course indicator. Elevating the antenna system higher to avoid local reflections destroys the course by introducing low-angle nulls in the vertical pattern of the system. The conventional counterpoise (35-ft diameter) does not eliminate the difficulty. Horizontal polarization, originally adopted for its superior performance in the instrument landing system localizer, aids materially in reducing the reflection amplitude. Probably this is because of its zero radiation at the horizontal angle.

Single reflecting sources at a site generally give smooth (sinusoidal) deviations of the course indicator. From a knowledge of the wavelength of the bends and the position where observed, the direction of the reflection source can be determined approximately by calculation. Distant hills, unless extremely elevated above the station and observed by a low-flying airplane, do not cause course difficulties. Good siting is important. Station sites must be fairly flat and free of the above reflecting sources for a radius of about 500 ft. The site should also be high because propagation at these frequencies is line-of-sight (about 45 miles is realized in an airplane flying 1000 ft above the station elevation).

RADIO MARKERS (75 Mc). An essential part of navigation is position checking along the route. The radio range gives only direction, or lateral, guidance. The null, or "cone" of silence, over each range station (Fig. 7) has been used for years as a means of determining "over the range" position. Intersecting radio range legs exist at some places and are used as "fixes" during flight. Special non-directional, low-power, low-frequency markers have also been used. The difficulties experienced with all of these may be summarized as follows: (1) unexpected fade-out of signal may be falsely interpreted as the zone over the station; (2) intersecting range legs are not always received in bad weather; (3) low-frequency markers have been heard many miles from their desired area at night. The development of the positive cone or *Z* marker overcame these difficulties.

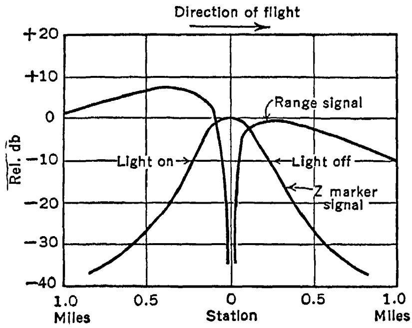


FIG. 7. Range Station Cone of Silence and 75-Mc Marker Signal Levels Experienced in Flight over Range Station at 1000-ft Elevation

The *Z* marker consists of a crystal-controlled, 3000-cps tone-modulated, 75-Mc, 5-watt, dual-transmitter and an antenna array. The transmitter is located within the range station and is a dual unit with automatic transfer in case of any failure. The antenna is located between the radio range towers and fed from the main station building by transmission line. The array may be considered as two pairs of horizontal, collinear, half-wave elements, at 90° to each other, and excited in phase quadrature. The array is one-quarter wavelength above a mesh counterpoise. The resulting service from this marker, when received by a longitudinal doublet under the airplane, is a circular pattern (Fig. 4) several times larger than the cone of silence and extending upward above the station about 10,000 ft.

The receiver generally used with these markers is a crystal-controlled, single-frequency, superheterodyne unit weighing approximately 25 lb and having a maximum available sensitivity of about 150 microvolts. Its audio output is filtered and connected so as to cause lighting of a signal lamp on the airplane instrument panel when marker signal is received. The audio output is connected to the pilot's headphone circuit. Three different kinds of indicator circuits are in use at present as shown in Fig. 8. The simplest is that employing a rectifier and relay, used by the AAF. One white light is connected to the relay contacts. The CAA and airlines use a receiver with no relays but with three output

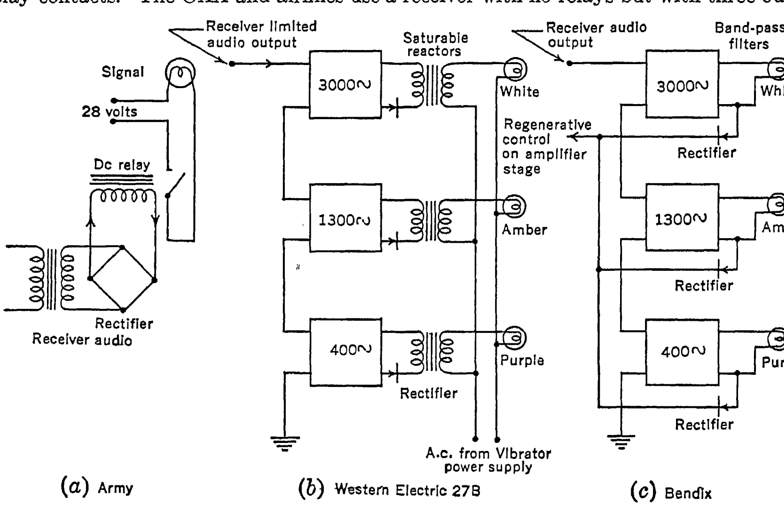


FIG. 8. Marker Receiver Signal Circuits

audio filters and three different-colored signal lamps. The white light, operating from 3000-cycle audio, serves for the airways markers (fan and *Z*). Amber and blue lamps, operating from 1300 and 400 cycles respectively, are used on the instrument landing system markers.

Good audio selectivity is achieved by designing the indicator-reactor circuit to operate at about 0 db and by limiting the receiver audio output electronically to about plus 6 db.

Any undesired signal, being able to produce only this limited output, will be inadequate to operate the lights if attenuated merely 6 db. Only simple filters are required.

The third system uses the audio signal directly to light the signal light. Good filters are employed. When any desired signal appears at the output, some of it is rectified and fed back to increase the gain of a controlled amplifier stage. This action gives regenerative response resembling relay action.

Horizontal half-wave doublet antennas are generally used for receiving marker signals. They are fastened longitudinally about 6 to 10 in. under the belly of the airplane. Sometimes the same antenna is also coupled to the low-frequency range receiver so as to serve a dual purpose. Some success has been achieved using a shortened antenna flush with the airplane skin and having a reflecting sheet inside.

Fan marker transmitting stations are keyed, tone-modulated, 100-watt, 75-Mc, dual transmitters connected by transmission line to a collinear dipole antenna array above a mesh counterpoise. Three thousand cycle tone modulation is used, and it is keyed in groups of dashes to identify the leg of the range on which it is located. The station is located on or near the center of the range course, usually about 20 miles from the range station (Fig. 4). Its antenna is aligned parallel to the range course and is received with greatest efficiency as the airplane passes over or to either side.

The name "fan" is derived from the shape of the marker field pattern. It is fan-shaped, as shown in Fig. 9, extending perpendicularly across the airway so as to be received even

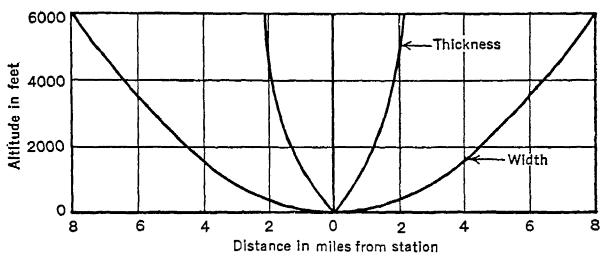


FIG. 9. Typical Dimensions of Airways Fan-type Marker

by aircraft considerably off-course. Its range increases with altitude but is generally receivable 6 miles off course at 1000 ft elevation.

Most of the 257 fan marker stations now operating have an antenna system using four collinear half-wave elements equally spaced 180° and carrying equal currents of the same relative phase. The array is one-quarter wavelength above a mesh counterpoise. Two minor lobes appear in the radiation pattern in the area directly above the station. A new array has been developed and is being installed in which the four elements carry currents in the ratio 1–3–3–1 and are physically spaced 220 electrical degrees. This arrangement eliminates the minor lobes and results in a dumbbell-shaped pattern in horizontal cross-section.

AUTOMATIC DIRECTION FINDER (ADF) RECEIVERS. The most popular and most useful radio receiver today is the ADF receiver (Fig. 2). It is an all-purpose receiver in the present airways system, providing for the reception of signals from radio ranges, CAA radio communication stations, airport control towers, and airline company offices. This receiver operates with a loop and a short-wire or vertical sense antenna. The loop gives accurate directional information as well as being very helpful in reducing the effects of precipitation and thunderstorm static.

The receiver operates as an ADF by virtue of the introduction of an audio modulation into the loop antenna RF circuit as shown in Fig. 10. The audio frequency is non-critical and is usually generated in the receiver at about 48 cps. The loop signal, with its locally superimposed modulation, is coupled through an RF transformer to the non-directional antenna circuit and to the IF amplifier circuit.

When the loop axis coincides with the direction of arrival of the radio waves, the loop contributes no RF signal to the receiver system and consequently none of the 48-cps signal gets through to the receiver IF amplifier or audio output. When the loop is turned so as to admit signal, the polarity (relative phase) of the 48-cps audio output with respect to the original 48-cps oscillator output is dependent upon the direction in which the loop is turned from its original null. A pair of thyratron tubes in a balanced modulator circuit compares these phases and through saturable reactors in the loop motor supply circuit drives the loop back toward the null. The motor and control circuit causes the loop to "seek" and hold the null position of any station tuned in on the receiver. A selsyn follow-up system is used to indicate remotely to the pilot the position of the loop. The loop position indicator is calibrated in degrees (0 to 360) and is called the bearing indicator.

In most installations there is some distortion of the received waves because of the structure of the airplane. This causes an error in indicated bearing called "quadrantal error." The error is of fixed amount in given directions, and the amount can be determined by experiment. Cam systems on the loop shaft are usually employed to compensate for the error automatically.

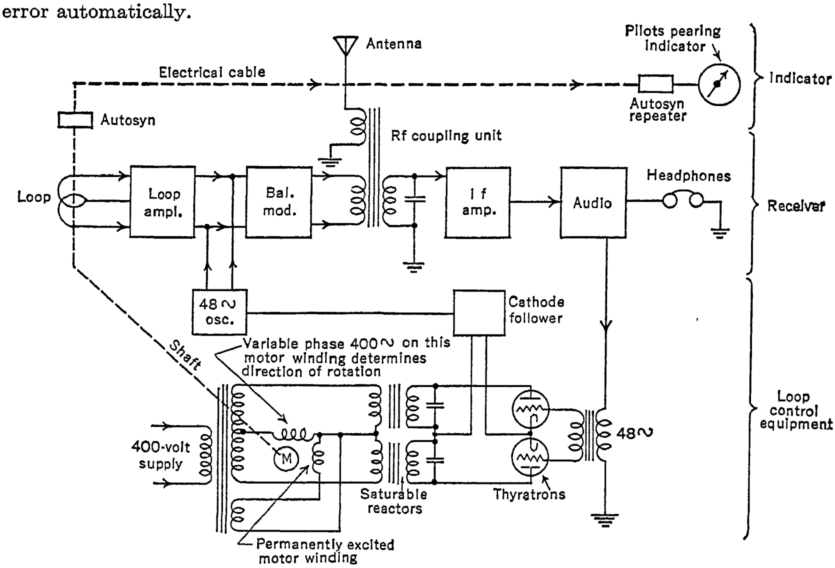


Fig. 10. Diagram of ADF Receiver Operation

COMMUNICATIONS (HF AND VHF). All communications in the United States airways system is by voice. The quality of voice signals received from aircraft, and sometimes that from ground stations, is far below any standard that would be acceptable in other services. As air traffic increases, there will be increasing need for quality, or for a different system, to preclude misidentification or misunderstanding of received calls. This condition is not easily cured. The major cause of poor quality is the enclosing structure needed around the microphone to exclude the 100-db noise in the present-day airplane cockpit. The ordinary microphone is overmodulated by this noise, and voice cannot be superimposed without severe distortion. The aircraft microphone is designed to be relatively insensitive by enclosing it in plastic. High-level voice, obtained by close talking, is admitted through small holes in the plastic.

Two other microphones are used. One, the "throat microphone," is designed to pick up vibrations from the throat by being worn on the neck. The other is the "lip mic" operating on the sound-velocity principle. Most of the noise pressures strike both sides of the armature of this mic and are canceled. By attaching it to the pilot's upper lip, high-level voice is applied to one side of the mic, causing modulation.

Service	Approximate Frequency Band
Long-range navigation	70–200 kc
Distress	500, 3105 kc, 121.5 Mc
Localizers	108–112 Mc
Radio ranges	112–118 Mc
Air traffic control	118–122 Mc
Airline communications	122–127 Mc
Air to ground	127–132 Mc
Glide path	329–335 Mc
Distance measuring	960–1215 Mc
Radar surveillance	3000 Mc band
Precision Radar	9090 Mc

Fig. 11. Chart of Proposed Frequencies for Air Navigation and Communication

Airborne transmitters generally operate in the 3 to 6 Mc band and have an output of 5 to 100 watts. The lower-power units are used by itinerant fliers in contacting traffic control towers. Plans are under way to convert all communications to VHF. Many ground station equipments at airport towers and airway stations are already installed. The VHF bands as outlined in Fig. 11 will permit numerous channels of static-free service. The aircraft antenna at VHF is more efficient than the short wires currently used, and consequently reliable service (not beyond line of sight) can be expected with the same or less transmitter power. All airline and itinerant transmitters are now crystal controlled. This practice will continue on VHF to insure reliable service.

AIR TRAFFIC CONTROL AND WEATHER REPORTING. The air traffic control system operated by CAA on the federal airways depends solely on voice communication between the controller and the pilot, assisted by fan markers and "holding" markers about which planes orbit until cleared for landing.

Advanced weather information is, of course, essential. Numerous weather stations now release radio-equipped balloons (radiosondes) to permit study of the upper air regardless of visibility conditions. These radio balloons emit coded signals revealing altitude, humidity, and temperature up to about 50,000 ft altitude. The radio equipment is expendable but is protected by parachute in its fall to earth; some are picked up and mailed back for reward. Direction finder and radar tracking of radiosondes has permitted determination of upper-air velocity and direction.

Two new electronic devices permit automatic measurement and recording of cloud-ceiling height and horizontal visibility, the "Ceilometer" and "Transmissometer" respectively.

The Ceilometer transmitter uses an extremely sharp vertical beam of high-intensity pulsing light. The source is a 900-watt mercury arc, striking 120 times per second. At a short horizontal distance from the transmitter, a photoelectric cell scans the entire beam from its base to top. If the beam is striking a cloud layer, the reflection will appear in the cell output. The vertical angle of cell at the time the output is observed is indicative of the height of the cloud layer. A recording device makes a record of the ceiling altitudes. A unique feature of Ceilometer is the use of pulsing light to eliminate the otherwise obliterating effect of daylight. Daylight, being steady, is filtered out of the system, permitting equally high ceiling measurement performance in daytime as at night.

The Transmissometer, which measures the light transmissivity of the air, is similar to the Ceilometer in utilizing a concentrated beam of light and a photoelectric cell. In the Transmissometer, however, the light is steady and directed horizontally to the cell through a kilometer of air near the airport approach lane. The effect of daylight is removed by shielding and baffling and by proper choice of beam intensity. The output of the cell is converted into pulses, the lowest pulses corresponding to low transmissivity. In the weather bureau these pulses are converted into relative values of direct current for observation or recording of visibility.

4. FACILITIES IN THE NEW FEDERAL AIRWAYS SYSTEM

The new federal airways system, now being placed in operation, will utilize VHF radio for ranges, instrument landing, and communications. The adoption of VHF relieves the troubles of static, both atmospheric and precipitation types.

The new radio ranges will be omnidirectional to satisfy the need for more airways, better traffic control, and most particularly to give useful navigation information regardless of position from the range station.

Distance-measuring equipment is planned as an ultimate replacement for fan and Z-type radio markers. Operating with the omnirange it will provide the basic requirement for safe air navigation—accurate knowledge of position at all times.

Congestion of traffic at air terminals will be reduced by the VHF instrument landing system now installed and by closer coordination in air traffic control.

In the new system all navigational information will be received visually, that is, by meter-type presentation. Some use of cathode-ray tubes may result from development work now under way, especially for anti-collision and air traffic control. Whether the presentation is by meter or cathode-ray tube, some effects from reflection of signal by buildings, trees, or mountains near the ground station will exist. These may cause bends, multiples, or wiggles in the course as indicated by the meter. In radar systems, wherein all the intelligence is obtained by visual study of the cathode-ray tube, the operator can generally separate the main from the reflected signal visually. Pulse technique does not in itself eliminate the effects of reflections, but when displayed against a time base on a cathode-ray tube it permits a study of all signals received. The reflected signals are visibly displaced and diminished by the extra time required to travel their longer paths.

INSTRUMENT LANDING SYSTEM (CAA). Instrument approach, or landing by instrument guidance, is just now being put into practice in the United States. It comprises a runway localizer, a glide path, and three marker beacons.

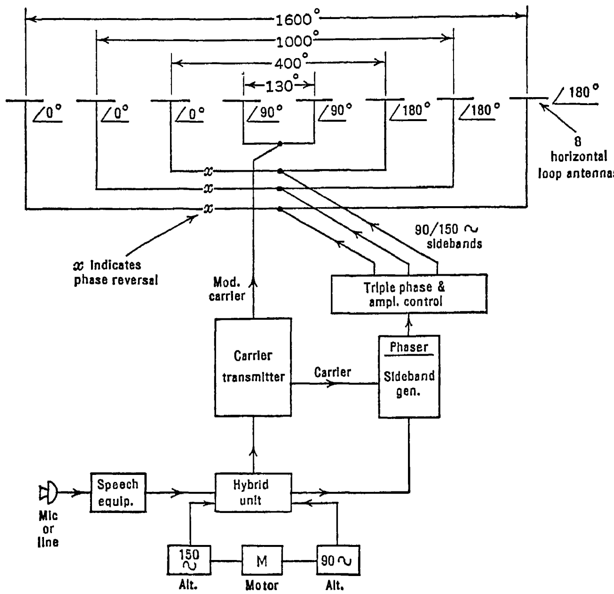


FIG. 12. Diagram of CAA Electronically Modulated Localizer

Localizer. The localizer creates a course, or track, along the center line of the runway, by overlapping two bean-shaped radio patterns having different modulation frequencies (90 and 150 cps). The signal service area extends slightly beyond the line of sight, but is conservatively given as 25 miles at 1000 ft (airplane elevation). The frequency is 110 Mc.

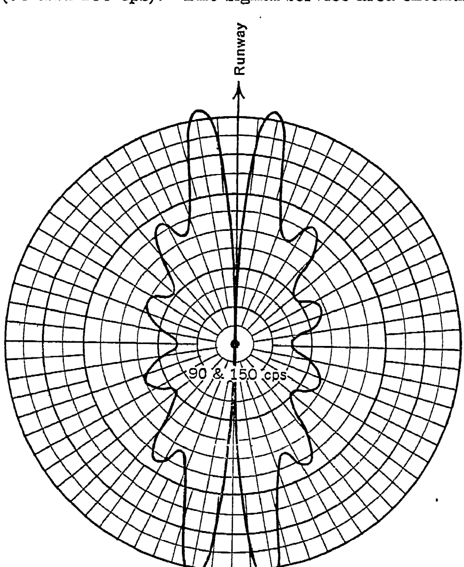


FIG. 13(a). Antenna Pattern of the CAA 8-loop Localizer. Side-band loops.

The localizer transmitting antenna array consists of eight horizontally polarized loops, spaced symmetrically across the center-line extension of the runway. It is generally 1500 ft from the end of the runway and elevated only about 12 ft. Each of the three outside loops on one side is paired electrically with a loop in similar position on the other side. Each pair is connected through a control to a common sideband generator, as shown in Fig. 12. There is a 180° phase reversal in the tie line between the pairs. The sideband energy is radiated with a sharp null down along the runway center line, as in Fig. 13*a*. The center two loops are connected in phase (Fig. 12) to radiate carrier, modulated with equal amounts of 90- and 150-cps voltage. The radiation pattern is shown in Fig. 13*b*. When the sidebands in patterns *a* and *b* of Fig. 13 combine in the receiver and appear at the filtered output, the overlapping patterns of Fig. 13*c* are obtained.

The airplane receiver, Fig. 14, is adjusted initially so that its filtered output of separate 90- and 150-cps voltage is well balanced. The output is connected to a balanced rectifier, and the rectifier is connected to a zero-center microammeter. The meter remains centered as long as equal amounts of 90- and 150-cps modulation are received, as when flying along the runway center line in the sideband antenna null. Deviation of the airplane right or left brings it out of the null. The sideband signal then received adds to that obtained from the carrier. On one side of the null (or course) the 90-cps signal adds to that of the carrier while the 150-cps signal subtracts. This causes deflection of the zero-center meter. The reverse is true on the opposite side of the course. The zero-center meter (Fig. 15) is used by the pilot to determine deviation from and direction to the course. The same instrument usually contains a second pointer, centered horizontally to indicate deviation from the glide path. The "U" receiving antenna is shown in Fig. 21*a*.

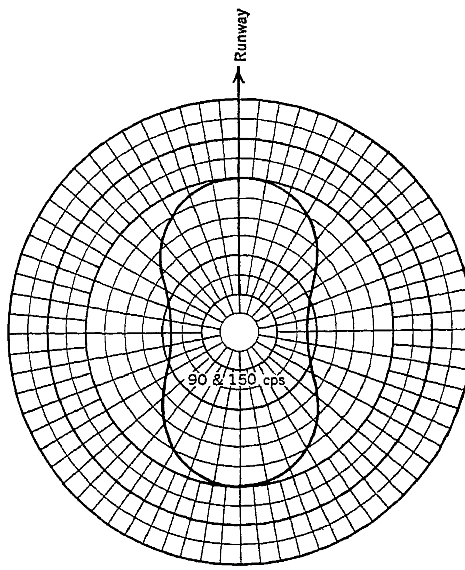


FIG. 13(*b*). Antenna Pattern of the CAA 8-loop Localizer. Carrier loops.

Although it is possible to obtain almost unlimited course sharpness in the localizer course by expanding the transmitting array and increasing the receiver output level, there is a maximum that can be used safely and conveniently by average pilots. The sharpness will probably be standardized at 5° (deflection to last dot left, or right, for 2.5° deviation from course).

Only part of the 200-watt carrier (40 per cent) radiated by the center pair of loops is modulated by the 90/150 cps signals. The balance is used to handle tone identification keying and control tower voice modulation. The latter is expected to be an aid in approach control of airport traffic.

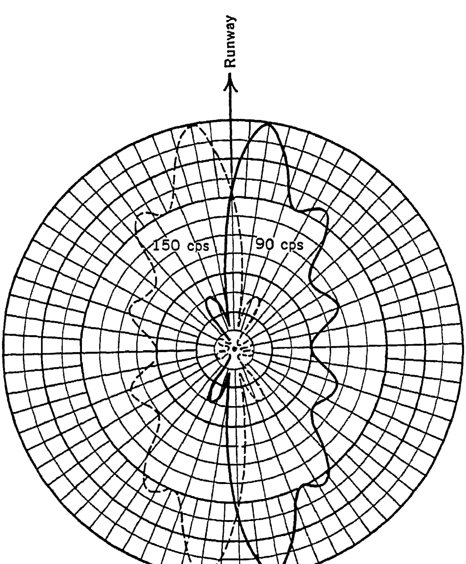


FIG. 13(*c*). Combined Side-band Pattern (*a*) and (*b*)

Glide Path. The glide path is established by a crystal-controlled transmitter capable of delivering about 25 watts at 330 Mc to an antenna system producing two types of lobes. One antenna produces a beam quite broad in the vertical plane (see Fig. 16*a*) which is modulated at 90 cycles. This will be called the upper beam. The other antenna is modulated at 150 cycles and is raised several wavelengths off the ground to produce a multilobed pattern with each lobe quite narrow in the vertical plane. Comparison of the broad upper lobe with the lowest of the narrow lobes establishes the glide path. This equisignal path is generally about 2.5°, although adjustment of the heights of the antennas can bring the crossover at any angle between 2° and 4°. An accurate plot of the lobes involved for the 2.5° path is shown in Fig. 17.

The separate modulation frequencies are generated by a mechanical modulator consisting of a synchronous (1800-rpm) motor and two metal paddle wheels having three and five paddles, respectively. The paddles detune associated resonant sections of transmission line coupled to the respective antennas and create 100 per cent modulation at 90 and 150 cps. The modulator is shown in the diagram, Fig. 16*b*.

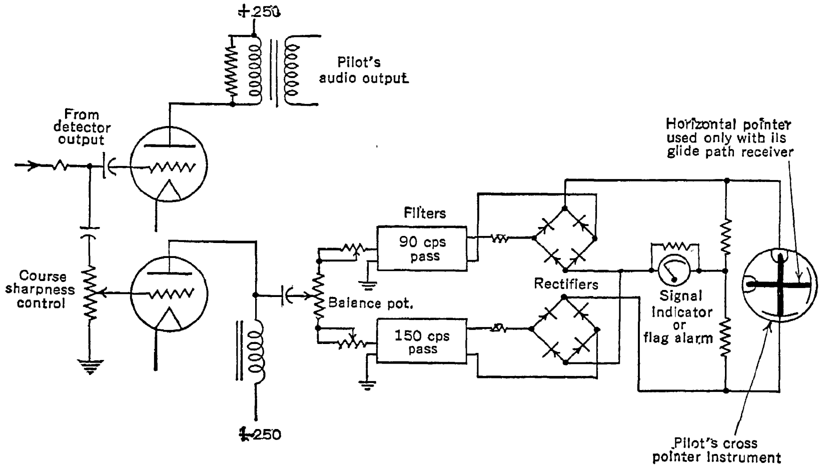


FIG. 14. Localizer Receiver Output Circuit

The equisignal surface represented by the two radiations of the station is obviously conical; paths from all directions terminate at the station. The station had to be placed off the side of the runway for safety reasons. This means that the approaching airplane, if in line with the runway, must follow a hyperbolic path whose minimum altitude exists directly opposite the station. To straighten out the bottom portion a special relation between the horizontal patterns of the upper and lower antennas had to be applied.

The receiver for the glide path uses a crystal-controlled superheterodyne circuit and a 28-volt d-c supply. It requires no high-voltage dynamotor or vibrator power supply.

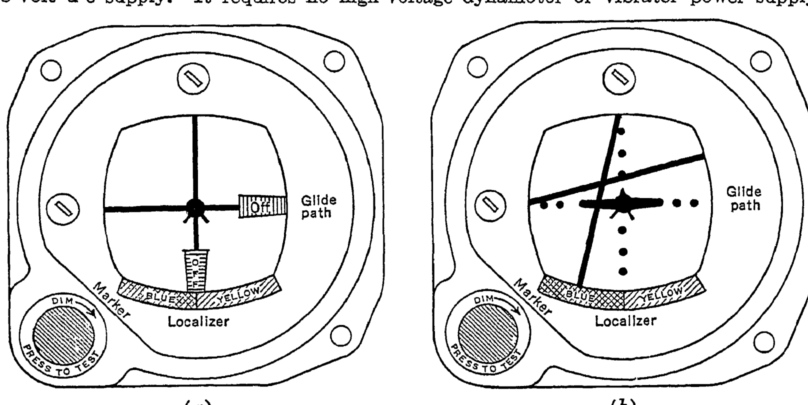


FIG. 15. Cross-pointer Instrument with Flag-alarms to Indicate Signal Failure. (*a*) Dead needle position showing alarm flags in position. (*b*) Operative position with alarm flags behind mask. (Courtesy AAF.)

The output audio (90 and 150 cps) is filtered and separately rectified as in the localizer receiver, Fig. 14. The resultant direct current is connected to a zero-center instrument (horizontal pointer of cross pointer instrument, Fig. 15). The pilot flies on the path, keeping the pointer centered, or horizontal. The dipole antenna, shown in Fig. 21*a*, is used in receiving the glide-path signals.

It is evident in that the glide-path equisignal lines converge and become very sensitive to vertical displacement of the airplane near the station and runway. This is offset by designing into the receiver about 9 db of negative AVC. This has the effect of reducing the receiver audio output voltage as the airplane approaches the station. An additional

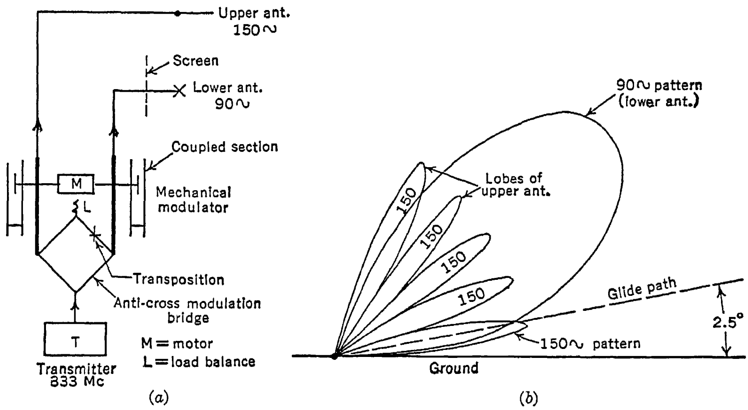


FIG. 16. 333-Mc Straight-line Equisignal Glide Path. (a) The lobes of elevated antennas produce an equisignal glide path. (b) Modulation is accomplished mechanically with coupled sections.

4-db reduction or "softening" of sharpness is derived by beaming unmodulated carrier across the path near its bottom end.

Markers. The markers used in instrument landing are, fundamentally, position-indicating devices similar to the Z markers discussed on p. 22-10. They will be replaced eventually with accurate distance-measuring devices which continuously indicate mileage to the runway. The present two markers are identified both aurally and visually by tone and keying. The outer marker (distance 4.5 miles) is modulated at 400 cps and keyed in long (2 per second) dashes. The middle marker (distance 3500 ft) is modulated

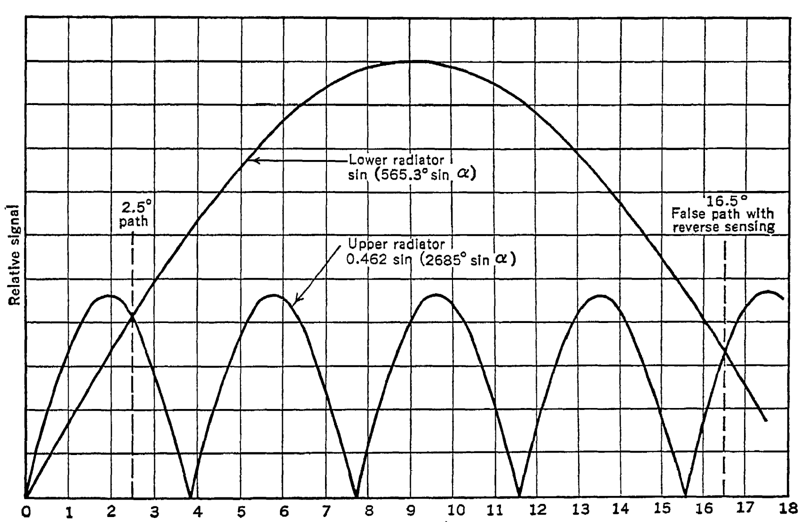


FIG. 17. Accurate Plot of Radiations Showing Formation of Glide Path

at 1300 cps and is keyed in a series of dot-dash characters. The boundary marker, recently discontinued, was modulated at 3000 cps and was keyed in fast dots (6 per second). The outer and middle markers are further identified by signal lamp color, purple and amber respectively.

THE OMNIDIRECTIONAL RANGE (VHF). This facility is produced by radiating simultaneously, from the same antenna array, two signals having the same audio-frequency modulation but having different relative phase for different azimuthal positions around the array. One signal is non-directional, and since its phase is everywhere (in azimuth) the same it is called the "reference phase" or voltage. The second is produced by rotating a "figure-of-eight" pattern. This pattern produces a modulation in the radio receiver, the frequency being dependent on the speed of rotation. Since its relative phase varies with azimuth it is called the "variable phase" or voltage. Since the two frequencies generated are equal, and a standard relation is set for north, comparison of the relative phase anywhere will permit determination of azimuth bearing from the station.

A five-loop transmitting array is used, and it is mounted on a tower 15 ft high with a circular counterpoise. Four loops are installed as diagonally opposite pairs. These are connected (Fig. 18) to the transmitter through a capacitance goniometer whose rotation

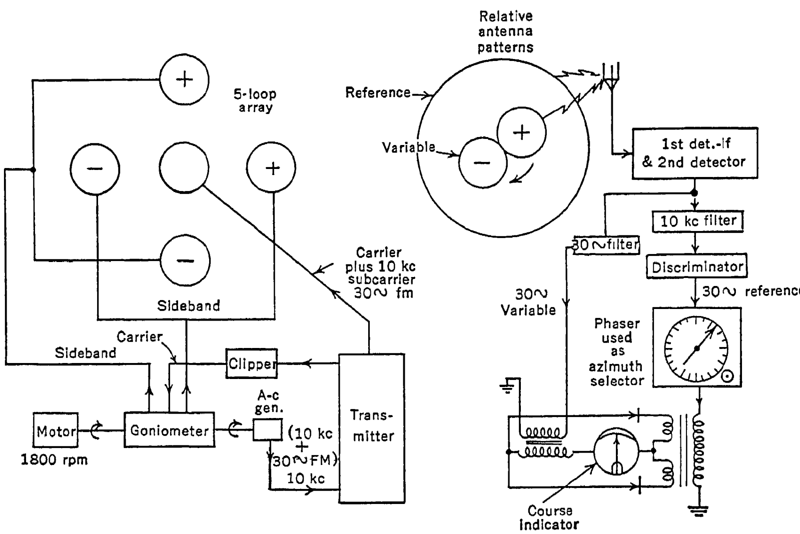


FIG. 18. The CAA VHF Omnidirectional Range System Diagram

at 1800 rpm causes the figure-of-eight space pattern of the loops to rotate in synchronism. It produces a 30-cps modulation of the carrier in the receiver.

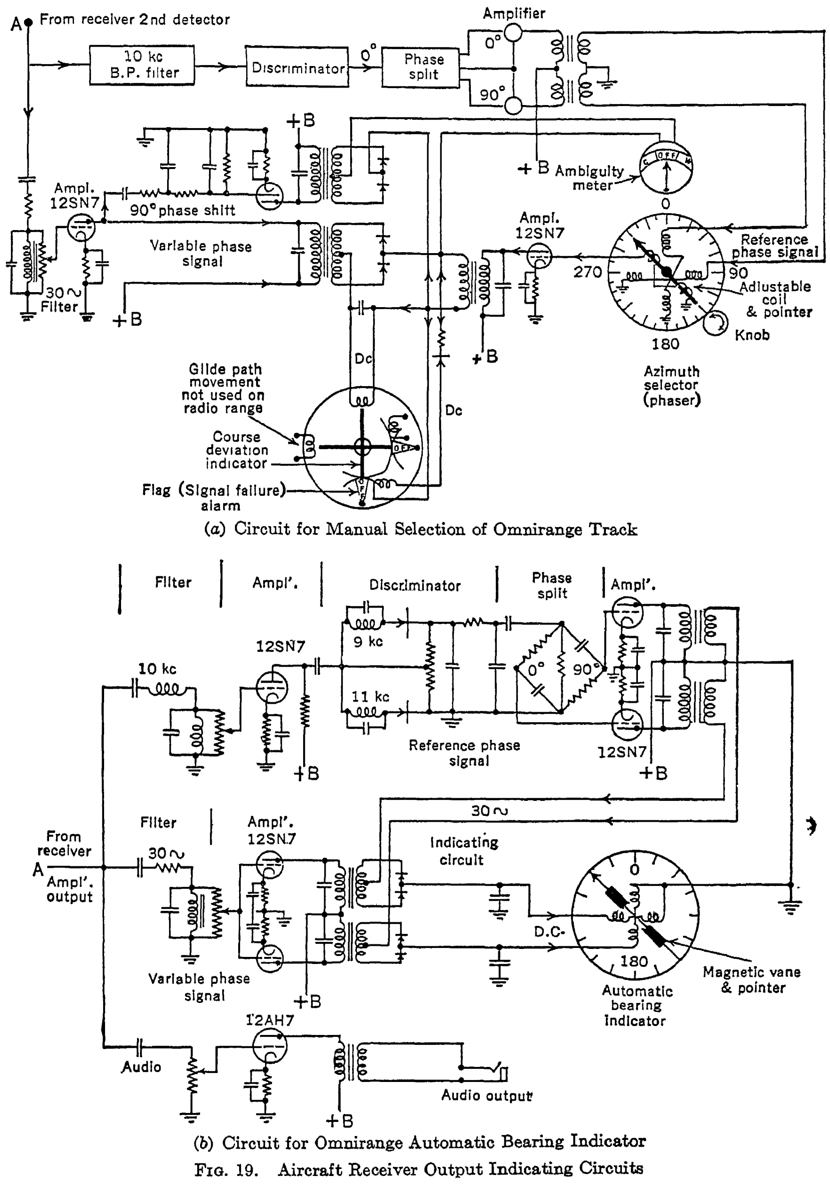
The center antenna is connected directly to the transmitter. The transmitter is modulated by a 30-cps voltage derived from a tone-wheel generator locked on the goniometer shaft. If the two 30-cps signals in the receiver output are in phase when the plane is north of the station, the phase difference in degrees elsewhere agrees directly with the degrees in azimuth of the receiver.

Two 30-cps signals cannot be directly isolated in the transmitter or receiver, so a subcarrier of 10 kc must be used for the reference signal. Further isolation is provided by frequency-modulating the subcarrier with the 30-cps voltage. Both the subcarrier and its required FM are derived from the tone-wheel pickup by using non-uniform spacing of teeth on the wheel. A "clipper" (see Fig. 18) is used to remove modulation on the RF to the goniometer.

The rotating pattern, which is sideband energy, effectively modulates the carrier about 30 per cent. The subcarrier modulation is equivalent to 10 per cent. The remainder of carrier is utilized for tone-keyed identification and voice communications.

The receiver used is a conventional superheterodyne up to the second detector. Here a filter is imposed to separate the 10-kc subcarrier and the variable 30-cps modulation. The 10-kc subcarrier is fed to a discriminator from which the reference 30-cps signal is obtained as illustrated in Fig. 18.

The two 30-cps receiver output signals are connected to a wattmeter circuit. The indicating element of the wattmeter is a zero-center instrument which is used as a course indicator. When the phase difference between the reference and variable signals delivered to the circuit is 90°, the indicating instrument will be centered. This may be standardized as the north, or east, bearings of the station. Varying of the receiving point from this bearing will cause movement of the zero-center meter. Full-scale reading is obtained in the system when the receiving point is displaced 10° from north. A variable phaser, Fig. 18, is provided to delay, or advance, the one signal before delivery to the wattmeter



(*a*) Circuit for Manual Selection of Omnirange Track

(*b*) Circuit for Omnirange Automatic Bearing Indicator

FIG. 19. Aircraft Receiver Output Indicating Circuits

circuit. If sufficient phase change is inserted to center the pointer of the indicating instrument, the amount of phase change inserted is proportional to the azimuth bearing. The phaser is calibrated 0° to 360°. This phaser setting is taken as the airplane-to-station bearing when the zero-center meter pointer is centered.

Obviously, there are two positions of the 360° phaser where the indicating instrument will be centered. This ambiguity is indicated either by a combination of red and green lights or by a smaller zero-center ambiguity meter. When signal lights are used, they are operated from a relay which is normally closed for correct signal and open for incorrect, or insufficient, signal. The relay closing is achieved by combining the two voltages in phase through an amplifier. Since the voltages are normally equal, reversal of one causes complete cancellation, the relay releases, and the red light comes on.

The alternative method uses a wattmeter circuit similar to that in the original course indication; however, its variable signal is shifted 90°. It, therefore, indicates a course 90° displaced to the true course. The ambiguity meter indicates full scale right until the airplane passes over the station. At this instant it swings to the opposite side, indicating that the station has been passed and that the bearing is reciprocal. When the ambiguity meter centers, it indicates failure of one or both signals or an inoperative receiver. One such meter is diagrammed in Fig. 19*a*.

Figure 19*b* illustrates an automatic "bearing indicator" for the omnirange. Its pointer follows up the bearing changes made by the airplane, always displaying correct bearing. It is a magnetic device with crossed coils connected to the respective rectified variable and reference phases from the receiver. Cathode-ray tube and other types of indicators have been developed. One system combines the omni- and the gyrosyn-magnetic compass to give the equivalent of VHF ADF.

DISTANCE-MEASURING EQUIPMENT. The distance-measuring equipment with meter indication depends for its operation on the challenging of a ground "radar transpondor beacon" by an airborne challenger. The challenger is a radio transmitter-receiver which transmits a pulse-type challenge to and receives a pulse-type reply from an automatic radio receiver-transmitter known as a transpondor beacon. The time elapsed between transmission of the challenge and receipt of the reply is a measure of the distance between the challenger and the beacon. This time difference may be converted to a d-c voltage whose magnitude is proportional to distance and can be read on a meter.

The challenge signal consists of a "pair" of RF pulses in the 960–990 Mc band so that a beacon setting must agree with two pulse characteristics before the beacon replies. These are the challenge frequency and the time separation (called mode) between the two pulses of a pair. This "frequency-mode" combination is called a challenge channel. The reply channel consists of a "pair" of RF pulses in the 1185–1215 Mc band. Challenge and reply channels in combination form an "operating channel." The beacon and the challenger receiver distinguish between pulses intended for them and pulses at the same frequency but different modes by use of a "decoder." One form of decoding is to delay the first pulse by an amount equal to the mode and operate only if the delayed pulse is then coincident with the second pulse.

An individual plane identifies the replies to its own challenges, as distinguished from replies of the same beacon to the challenges of other planes, by the fact that its own replies occur after a fixed (or slowly changing) time (i.e., distance) after its challenge, while the replies to other planes occur at a random time. The airborne receiver is "gated on" for a very short time interval at a fixed time (i.e., distance) after each challenge. It is as if the receiver asked the question, "Is there a reply from a beacon at a distance between 20 and 22 miles?" The gate duration is defined as 22 minus 20 or 2 miles. In one design the gate is divided into two subgates called (1) the "early gate" and (2) the "late gate." If the signal occurs in the early gate, the time delay of the gate is decreased for later challenges (it is gated on at 19 instead of 20 miles). Likewise, if the signal occurs in the late gate, the time delay is increased for later challenges (it is gated on at 21 instead of 20 miles). This system of subgates allows automatic following of the signal. The time between the challenge and the "gate" is translated into volts by a measuring sawtooth voltage which defines the volt-time (i.e., volt-distance) relationship.

While searching for a beacon, the gate is made to travel slowly over the total distance, advancing a fraction of its duration after each challenge. It is as if the gate asked the question, "Is there a reply between 20 and 22 miles?" then after the next challenge it asked, "Is there a reply between 20.5 and 22.5 miles?" etc. If the gate finds a high enough percentage of replies to challenges, the search is terminated and tracking begins; that is, the receiver thereafter holds the signal which has been found.

Each beacon is identified by "gap coding"; i.e., its transmission is interrupted by a keyer in such a manner that the gaps form a Morse character. Failure to receive a reply results in lighting a lamp in the cockpit which therefore flashes the beacon's code.

A beacon is disabled for a short time after replying. This time is called "beacon recovery time" or "beacon dead time." While this "recovery time" may serve many useful purposes such as preventing over-interrogation, it also results in failure to reply to those challenges which occur while the beacon is recovering from a previous interrogation. This failure is called "count-down" and is expressed as the ratio of missed replies to the total number of challenges.

The number of aircraft signals that can be handled simultaneously by a given beacon is very important. The system described above, through choice of pulse dimensions and random rate, can handle approximately fifty airplanes per channel simultaneously (sixty channels are repeated in every 500 mile square). A block diagram of the airborne unit is shown in Fig. 20.

Distance can also be measured by the same principle as that employed in the radio altimeter, but with the aid of ground respondor stations. Continuous-wave frequency-modulated transmission is employed. The modulation is varied over a cycle which is, in duration, equal to at least the time required for travel of the signal to and from the most distant beacon to be measured. The frequency of the received signal at any instant is different from that then being emitted. The difference is a function of the time and therefore the distance. The meter-style indicator can be calibrated in miles. The disadvantage of this CW system is that it can handle only one aircraft at a time.

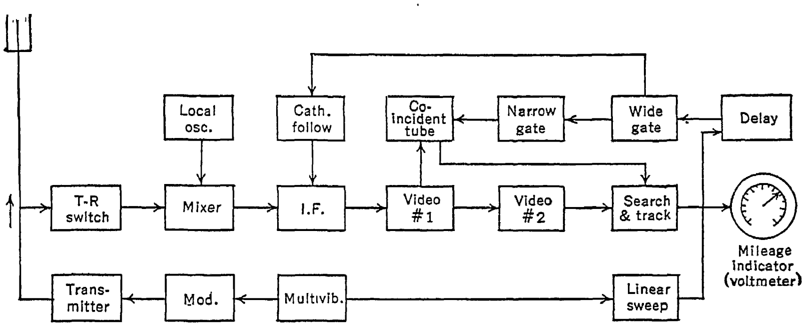


FIG. 20. Functional Diagram of Distance-measuring Equipment (Airborne Unit)

Audio phase comparison methods have been proposed and tried experimentally. In this method an audio modulation frequency is chosen whose wave period (360°) is the same as the time required for transmission to and from the maximum distance to be measured. The radio signal travels 186 miles in 1 ms. Therefore, a 1000-cps frequency will shift 360° in 186 miles. If the airplane and ground station are 93 miles apart, the round trip between them is 186 miles and the 360° shift will prevail as the maximum range for the 1000-cps wave.

By sharing time through random keying of the airborne transmitters, several aircraft may obtain separate distance information simultaneously in the audio phase comparison system from the same ground station.

AIRCRAFT EQUIPMENT (RECEIVING). Almost universally, aircraft receivers have been of the superheterodyne type. Tunable receivers have been used generally, with spot frequency for control towers (278 kc) crystal-controlled or fix-tuned. In the VHF system the superheterodyne principle is again being universally used. Crystal control is being used for most commercial applications and manual tuning for inexpensive units for itinerant fliers. In the crystal-controlled receivers each of the several crystals employed will produce a multiplicity of receiving channels. Over two hundred channels are needed by aircraft on routine instrument flights, in the bands illustrated in the national plan, Fig. 11.

Many companies have, or are installing, dual ADF receiving equipments in each airplane so that automatic indication, or plotting, of aircraft position can be used. Single instruments with dual pointers, one for each ADF receiver, and each receiver tuned to a separate station, are very helpful in navigation.

VHF receivers for navigation are equipped to serve all VHF functions required in flight. Two such receivers in an airplane give adequate stand-by protection. In normal use one is available for communication while the other is being used for navigation or landing.

Two types of output indicating circuits are now used in each of these navigation receivers. The amplitude comparison type (90–150 cps) will be retained only until existing localizers of the instrument landing systems are converted from amplitude to phase comparison principle. Then the phase comparison type indicating circuit will be used for both omni navigation and landing.

An outstanding advancement in receivers was made during the war in the elimination of all receiver high-voltage power supply. Through the development and application of the 28-volt-type tube (28D7) a complete 330-Mc, superheterodyne, glide path receiver, using no vibrator or dynamotor-type supply, was produced in quantity. Twenty-eight volts direct current, available in the airplane, is used as plate supply. An ADF receiver was developed later, also avoiding the HV supply. The reduction in noise and maintenance and the saving in weight by this development are extremely important factors.

Receiving antennas have been an important part of the VHF program. Whip or mast-type vertical antennas are the best for ease of installation. For lateral guidance functions, however, horizontal polarization has been declared superior. Suitable antennas are illustrated in Fig. 21. The patterns of the U and V antennas are essentially circular, as required for their navigational function in range and localizer receptions. The 330-Mc dipole on the U antenna, in Fig. 21, is used for glide path reception. It is required to have only forward glide path reception during the runway approach procedure.

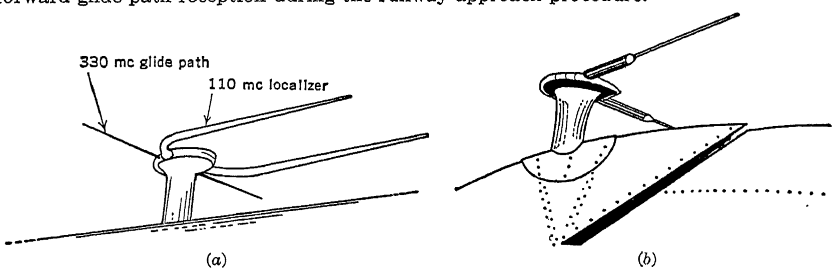


Fig. 21. Receiving Antenna for the Instrument Landing System. (*a*) The U is for 110-Mc band localizers, and the dipole is for the 330-Mc glide path. (*b*) An experimental V antenna for non-directional reception of localizer or range signals.

For navigational service, pure polarization is essential in the radiated wave. If polarization is impure, that is, unintentionally mixed, serious attitude effects are evident in the indicator when the airplane is banked or turned.

Balanced transmitting and receiving antennas are now used.

AUTOMATIC FLIGHT AND LANDING EQUIPMENT. For automatic navigation, the new electric auto-pilot has become the greatest asset. With it, straight and level flight is made by gyroelectric control, or completely coordinated (aerodynamically) turns of the airplane may be made by turning a knob which simply unbalances electrical bridges in the auto-pilot system. If the bridge circuit is electrically connected to the output of the new navigation receiver (localizer or radio range) through an appropriate amplifier or coupling system, the auto-pilot can be made to fly the airplane accurately along the localizer or radio range course. Or the output of the receiver (direct current which is proportional to displacement from the course and whose polarity reverses when crossing the course) may be made to operate a steering motor right or left, to follow the course. The steering motor turns the auto-pilot steering potentiometer.

The operation of one type of electrical auto-pilot, and one way in which the radio signal may be coupled to the auto-pilot, are shown diagrammatically in Fig. 22.

Coupling the radio guiding signal to the auto-pilot and obtaining satisfactory performance involves consideration of the mass and speed of the airplane, sharpness of the radio course, and characteristics of the auto-pilot itself. The desired performance is that giving asymptotic approach to the course. In the off-course position, a "displacement" signal must be applied to turn the airplane right or left as required. The displacement signal is obtained directly from the radio receiver and is proportional to the angular distance from the course. Acting alone, this displacement signal would reduce the turn of the airplane to zero as it crosses the course. But the airplane heading at the time of crossing may be at any angle to the desired course. The airplane travels to the opposite side of the course before a reverse signal is applied. The result of displacement signal alone would be a continuous oscillation of the airplane across the course as it flies toward the guiding radio station.

In the coupling device, a "rate" signal must be applied—a signal whose amplitude is proportional to the rate of change of displacement, and whose control on the airplane through the auto-pilot is reverse from that of the displacement signal. With proper design for any particular airplane, the rate circuit reverses the turn of the airplane as it nears the course and causes it to follow the desired asymptotic curve. Because of the converging sharpness feature of radio courses, a compromise must be made in the coupling system characteristics. In general, however, good performance has been achieved. On the final approach to the airport on the localizer, the vertical pointer of the course indicator seldom deviates from center by more than its own width, regardless of any cross-wind velocity or direction.

Flight toward or away from the station is obtained by reversal of connections (polarity) between the radio receiver and coupling unit. The switch positions are marked "inbound" and "outbound" respectively.

Very flexible radio range navigation is now possible through the use of computers. For the federal airways system these computers would operate in conjunction with the VHF omnidirectional range and distance-measuring equipment. With the computer, it is

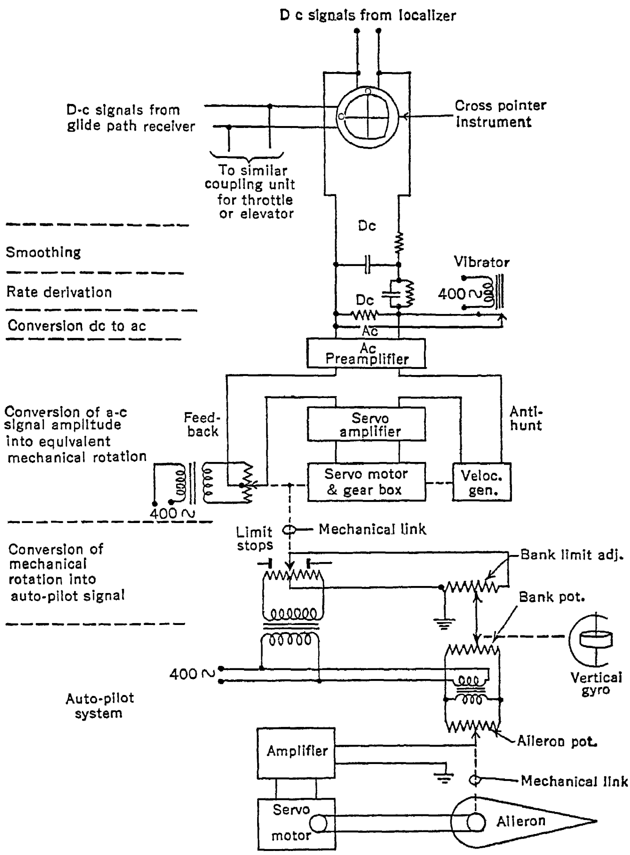


FIG. 22. Auto-pilot and Radio Coupling Systems

possible to fly a synthetic left-right indicated course in any direction, whereas the regular courses are defined only toward or away (radially) from the station.

Referring to Fig. 23, S represents the omnirange station, and the course to be flown lies along the non-radial line AB. The distance-measuring equipment (described above) produces a voltage e_1 which is proportional to r, the distance to the airplane. This is applied to a selsyn primary (rotor) which is connected to the omnirange azimuth indicator. The sinusoidal output of the selsyn secondary winding is

$$e_2 = e_1 \sin \theta$$

because it is designed to vary sinusoidally with rotor angular displacement. Since e_1 is proportional to r, the equation may be written:

$$e_2 = kr \sin \theta = ky$$

The voltage e_2 therefore is proportional to the variation in y, that is to the displacement of the airplane from the line AB. This displacement voltage may be balanced by a fixed voltage and presented on the usual left-right course indicator for manual flight, or it may be coupled to the auto-pilot for automatic flight along the selected line AB.

In setting the computer so that the constant k will be properly handled, the course line to be flown is decided upon and drawn on the range map. The direction and length of the perpendicular from the range station to this line are determined. The direction of the perpendicular is set into a calibrated clutch between the azimuth indicator and the selsyn unit. The length of the perpendicular line is set up on the "lane selector" switch, thereby providing the required fixed balancing voltage. Then the pilot's left-right indicator will be centered only when the airplane is on the desired flight line.

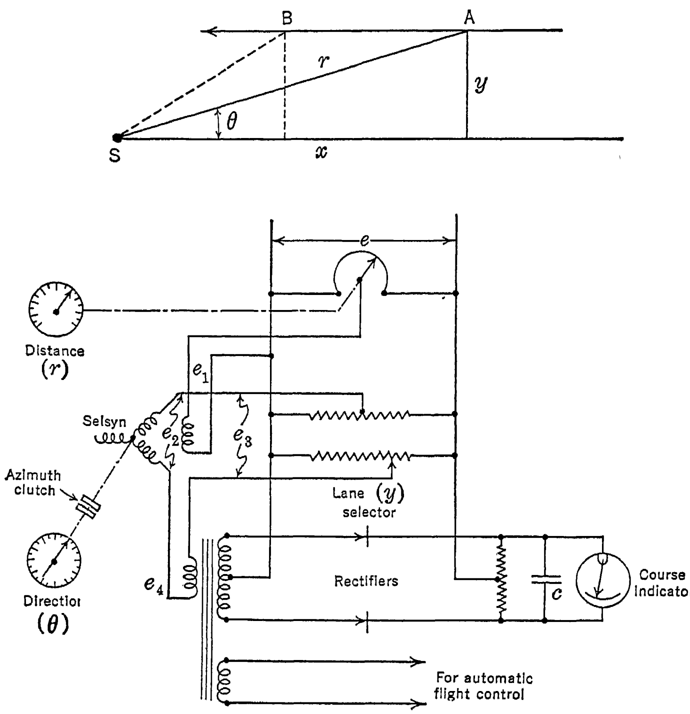


FIG. 23. Computer for Synthetic ($r\theta$) Courses

RADAR MONITOR FOR AIRPORT TRAFFIC CONTROL. Civilian aviation will require the use of radar as a monitor in airport control towers. Development of suitable equipments is under way. Its first use will probably be in expediting the outbound traffic at the congested terminals. The airport tower operator, having accurate knowledge of the displacement of all inbound airplanes, through reference to the radar monitor, may permit the departure of airplanes that otherwise would be required to await position reports from the inbound airplanes.

Three fundamental problems are evident in the application of radar to airport control towers. These problems are typical of all civil radar application. They were overcome in military use by a multiplicity of radar equipments and large operating crews. The problems are: (1) reliable coverage up to about 30 miles, from horizon to zenith; (2) the elimination of undesirable ground clutter; (3) the presentation of resulting radar information in a manner such that the regular tower crew can use it safely, at any instant day or night, without special enclosures.

Relatively high-angle coverage has been obtained by means of antennas giving cosecant-squared patterns, shown in Fig. 24. When the energy is distributed in the required pattern, the maximum distance range is naturally reduced. A peak power output of about 0.5 megawatt is required to give the 30-mile horizontal coverage with a cosecant-squared pattern.

Some work has been done on the elimination of ground clutter as illustrated in Fig. 25. Only moving targets are permitted to appear. Development is being continued to simplify the equipment and reduce the maintenance required to keep it in perfect adjustment.

The presentation feature has not yet been satisfactorily solved. New, high-intensity cathode-ray tubes are being made available as an aid to the daylight presentation problem. Another means of getting the radar scope picture visible in daylight is complementary light filters. For example, if the control tower window glass is colored blue and an amber filter is used over the face of the radar scope, the picture will appear the same as at night. The reflections of the observer's face can be eliminated by installing the amber filter at a 45° angle.

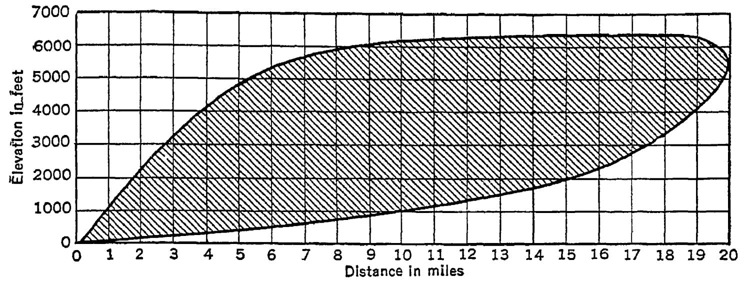


FIG. 24. Cosecant-squared Pattern of GCA Search Antenna

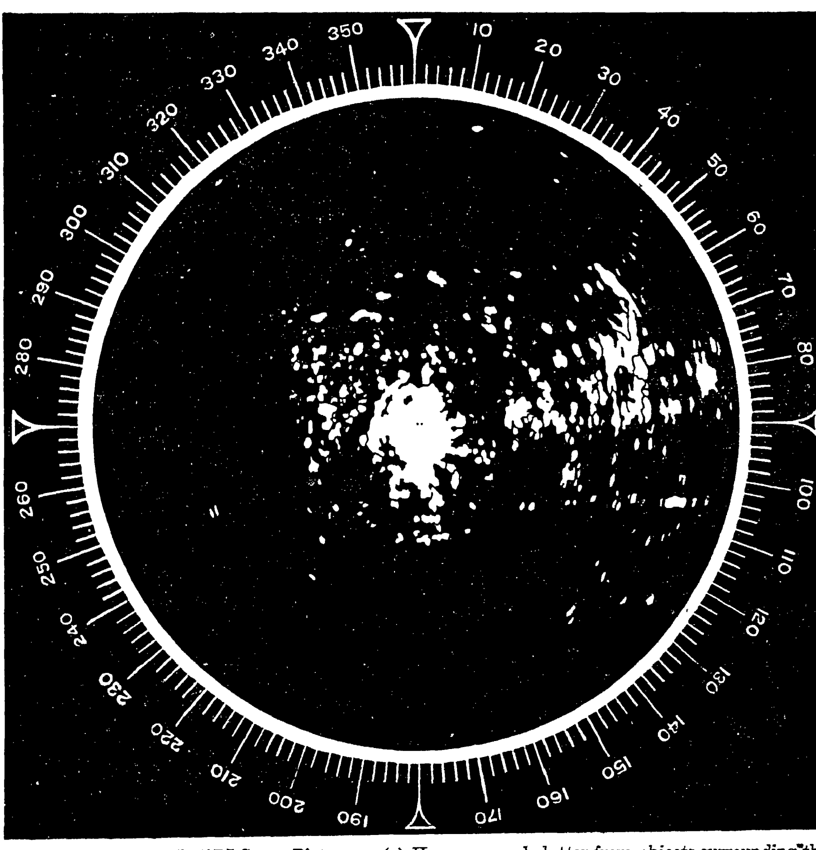


FIG. 25. Radar 10-mile PPI Scope Pictures. (a) Heavy ground clutter from objects surrounding the station at the Indianapolis Airport. The targets between 60° and 100° are buildings in Indianapolis.

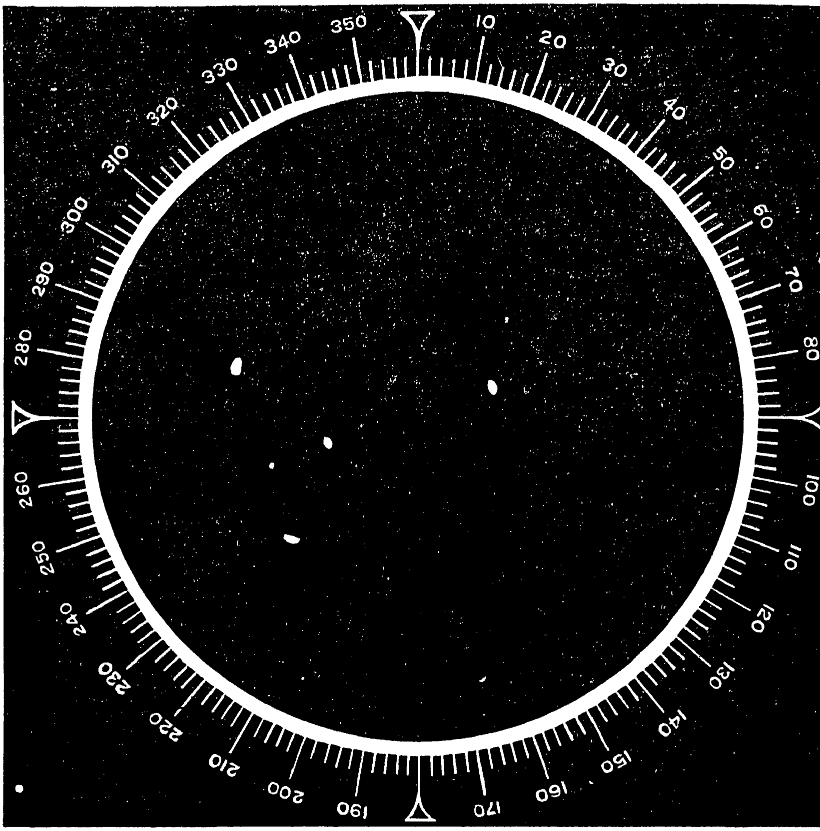


FIG. 25. (*b*) Same as (*a*) except that Ground Clutter Is Removed and Four Airplanes Are Shown in Plan Position (Courtesy CAA)

5. PROPOSED NEW LANDING SYSTEMS

GROUND CONTROLLED APPROACH (GCA). At the beginning of the war, a talk-down system was conceived for landing airplanes in bad weather. It became known as GCA because it served in reverse to other systems—it controlled from the ground. Several sets of GCA were put in service before the end of the war. Many sensational landings of military airplanes were made with GCA, saving lives and valuable property. Its use required no new airborne equipment but depended only upon communications and a cooperative pilot.

The GCA equipment is a composite, trailer-type station with three radar systems and complete communications equipment. It is placed about 500 ft to the side of the runway and at the end opposite from that on which the landing is to be made. Air-conditioning and power-generating equipment are carried on the towing truck.

Originally the GCA trailer required a crew of five operators. Four of them constantly watched four radar scopes, and the fifth served as final approach controller to give heading and rate-of-descent instructions to the pilot on the final approach. Now, however, the operation of landing a single airplane may be handled by one operator who shifts his attention from search to precision scopes as the airplane orients and approaches the runway, Fig. 26. If more than one airplane is involved, more operators are required.

The function of GCA, whether operated by one or more men, is as follows: (*a*) *search* for aircraft in all directions, using PPI scope presentation; (*b*) *direct* aircraft into the landing sector about 6 miles from the runway at 1500 ft elevation, on the basis of search radar information; (*c*) *land* the airplane by giving the pilot explicit instructions constantly during the approach as to heading and rate of descent, on the basis of precision azimuth and elevation radar information.

The search function in GCA is obtained at 10 cm by an antenna rotating at 30 rpm. The antenna is a special reflector with two-dipole array fed by rectangular wave guide. Its radiation pattern is illustrated in Fig. 24.

The precision radar system consists of one transmitter (3 cm) sharing time with two special antennas. One antenna is vertical and the other horizontal. Each antenna consists of a multiplicity of collinear dipoles mounted along a wave-guide section and fed from

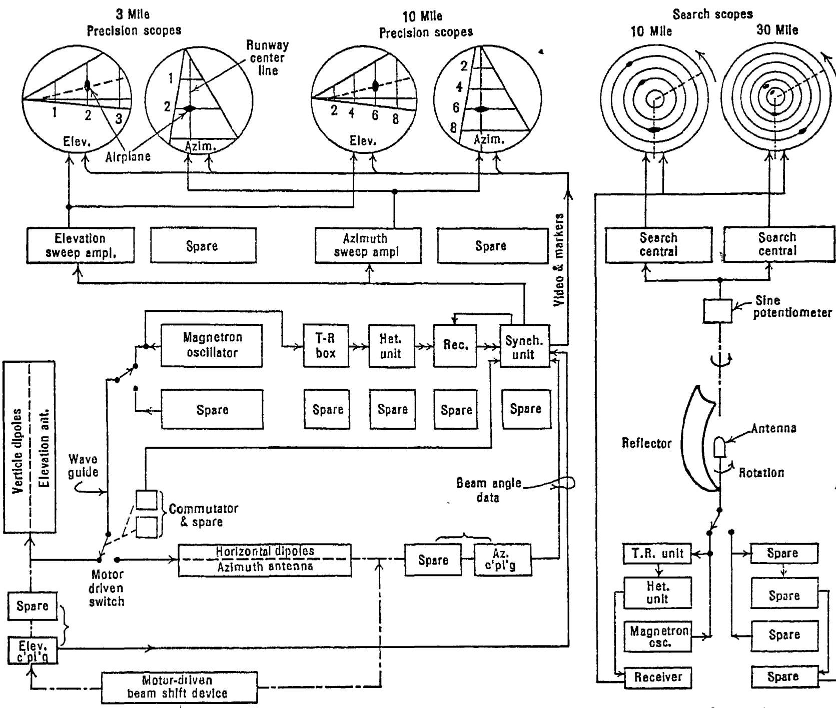


FIG. 26. GCA Search and Precision Radar Equipment Arranged for One-man Operation

small probes projecting into the wave guide. The spacing of the dipoles and phase of their currents create the sharp patterns required for precise direction determination. These patterns are aimed along the approach path and then made to oscillate rhythmically across the path by shifting the dipole current phases (mechanically distorting the rear of the wave guide through a motor-driven mechanism).

The received precision radar echoes are displayed on offset-center PPI scopes as illustrated in Fig. 26. This presentation and the narrow radiated beams permit the operator to obtain great precision in the observation of aircraft displacement. The vertical antenna gives the glide path displacement, and the horizontal antenna gives lateral displacement.

One unique feature in the GCA, which is used in certain other radar equipments, is the optical system of the PPI scopes. The operator views an illuminated map through a 45° glass plate. The scope is directed at the glass from a position complementary to the position of the map below. The scope appears superimposed on the map. It is on the map that the correct approach line is inscribed for use by the controller in detecting deviation of the airplane.

Three important features of GCA are: (1) No equipment is required in airplanes for its use other than already existing communications. All aircraft then may use it. (2) Its precision and straightness of path are unaffected by conditions surrounding the site, such as buildings, hangars, and wires. (3) The precision is not entirely in the system but partly in the ability of the operator to bisect the target image. (4) The identity of aircraft being controlled depends largely upon the skill and attention of the operators. (5) Skill, practice, and experience are required in any blind approach operation. In GCA this skill is on the ground.

NAVAGLIDE (Federal). The navaglide instrument landing system being developed by the Federal Telecommunications Laboratories is a microwave system using only one frequency channel. The four-directional signals of the glide path and localizer indications will operate on this frequency channel simultaneously through the use of a scheme of subcarrier modulation. The receiver of the "Navar" navigation system (described below) may be used, ultimately, also for the landing signals. Accurate distance-measuring equipment will replace the markers of the present instrument landing system, permitting use of automatic landing equipment.

MICROWAVE (Sperry). A complete microwave instrument landing system, operating at approximately 2600 Mc, has been developed and successfully operated experimentally for several years by the Sperry Gyroscope Company. This development, now in production status, overcomes objectionable siting and wave-reflection problems existing on systems using lower frequencies. Reasonably small radiating systems are able to concentrate and confine the radiated energy to the approach sector and avoid nearby buildings, wires, or hangars. Reflections from these objects on other instrument landing systems causes course and path bends.

The present success of the system may be attributed to the development of the klystron tube (2K36/416) for airborne receiver use and to the development of crystal control for both the transmitting and receiving equipment.

The complete system consists of two ground stations (localizer and glide path) and one airborne receiver. The ground stations are identical in electrical circuits and have equivalent outputs of about 70 watts but use different radiating elements. The glide path radiator is a tilted vertical parabola fed by two wave guides through a mechanical modulator. The wave-guide feeds are displaced from the parabola focus in opposite directions and therefore produce two overlapping patterns, the plane of overlap being inclined upward at the desired approach glide angle. The patterns are amplitude modulated 600 and 900 cycles respectively. The parabola is narrow to give wide horizontal coverage. This permits its being placed safely to the side of the runway.

The localizer transmitting antenna is a paraboloid, giving a relatively concentrated pattern in both horizontal and vertical planes. It is equipped with a central vertical dividing shield to separate the two wave-guide feeds. The radiated patterns overlap in a vertical plane thereby forming a course for left-right guidance along the runway center line. The localizer is placed on the runway center line at a safe distance from the end opposite from that on which the airplane lands.

The one receiver used in the airplane has a common crystal-controlled klystron tube oscillator serving two separate IF circuits. The filtered and rectified outputs of the two channels are balanced and connected to the localizer and glide path pointers respectively of the conventional crossed-pointer landing instrument. Two flag alarms on the instrument serve as indicators to warn against failure of the localizer or glide path signals.

An electronic coupling device now in production is used to couple the localizer and glide path signals into the Sperry (A-12) electric auto-pilot for automatic approach flight. It works equally well on the CAA instrument landing system.

6. PROPOSED NEW SHORT-RANGE NAVIGATION SYSTEMS

LANAC (Hazeltine). The word "Lanac" is derived from "*la*minar *n*avigation and *a*nti-*c*ollision."

This system is proposed for the 1000-Mc band and utilizes in each airplane an interrogator and a replier. The replier consists of a pulse receiver and transmitter capable of automatically replying to interrogating pulses of proper frequency and coding. The code key of the replier varies with altitude layers in a prescribed standard manner, through the use of an aneroid cell.

The interrogator includes a pulse transmitter, coded to challenge the repliers of other aircraft, and a receiver to interpret the reply. Normally it interrogates on the code equivalent to its own altitude so as to provide anticollision and traffic control safety. By means of a switch, the interrogation code can be varied to scan the traffic in altitude layers above or below the level being flown. The interrogator antenna is directional and rotatable so that direction as well as distance and altitude of replies can be observed on the L-type (or possibly PPI) scope used in the system.

Ground transpondors and ground interrogator stations are included to permit route navigation by the aircraft and ground surveillance of airways traffic overhead, obtaining position of planes in three dimensions as well as identity of each plane, and affording selective signaling to each plane. The ground transpondors also serve as obstruction warning units. Ground transpondors, properly placed on the approach, aid the airplane on instrument approach to the airport.

The Lanac interrogator may be employed as a radar, in special applications, by tuning the receiver and transmitter of the interrogator to the same frequency so that echoes will be shown on the interrogator's display in the usual radar manner. In this case, of course, no replier is utilized.

This radar mode of operation is used in an aircraft to supply terrain-clearance information when the plane is flying at an altitude of 500 ft or more. It therefore offers an important safety feature for off-airway flying over mountainous country. The radar mode is used in marine anticollision service to warn against unequipped craft. Radar anticollision protection between aircraft is not feasible because of the small size of the targets and their extremely high relative speeds when closing on a head-on-collision course.

The Lanac system is useful in marine navigation, and a description of this application is given under Marine Aids.

TELERAN (RCA) (*tele*vision *r*adar *a*ir *n*avigation). Teleran is a comprehensive system of navigation involving radar as the means of collecting air traffic information and television as a means of displaying the information to the pilot. Radar stations with overlapping 50-mile service areas to form the airway are proposed. Respondors in every airplane are essential, although failures of respondors can be taken care of in the plan by using separate echo-type search radar equipment. The PPI picture obtained in the ground radar is used by air traffic control or tower operators. This same picture, with any desired obstruction, control, or weather instructions superimposed, is televised, transmitted to, and repeated in the airplane by television.

The system permits both pilot and ground operator to see and appreciate the complete air traffic situation. The airborne transpondor units can be coded or varied automatically with altitude to segregate the various flight levels and to permit identification of aircraft. Instructions may be written out on the ground and "handed" to the pilot on the television picture.

For landing at an airport, an airport localizer radar and GCA precision radar units are proposed.

For each altitude level and for the echo-type search radar, separate television transmitting equipment and radio-frequency channels are required. The channels required can be greatly reduced over commercial television because a low scanning rate may be used.

NAVAR (Federal) (*nav*igational and traffic control rad*ar*). A system for traffic control and navigation around airports and along airways. This system provides the following features, which may be applied progressively to an airways system: (*a*) Ground radar surveillance in the form of PPI displays. (*b*) Distance and azimuth information in the airplane. The azimuth information is omnidirectional derived from the ground radar system. (*c*) PPI traffic presentation in the airplane, relayed from the ground station. (*d*) Selective signaling, ground to aircraft; and automatic identity and altitude response, aircraft to ground.

The ground radar equipment and display is conventional except that provision is made to separate known (respondor-equipped) aircraft from others. The distance information in the airplane is obtained by a pulse interrogating system. The azimuth information in the airplane is obtained by measuring the time from the reception of a non-directional pulse radiated from the ground to the reception of the rotating search radar beam. The pulse and beam are synchronized for true north direction. The timing measurement in the airplane is made automatically.

The airborne PPI display, called "Navascope," is obtained by sending synchronized pulses omnidirectionally from the ground radar station. These pulses contain the information of all aircraft in the area as revealed on the ground radar scope and are reproduced in synchronism on the airborne PPI. The airborne PPI display includes positive identification of observer's own airplane and self-centered altitude layer presentation.

Selective signaling of aircraft is obtained by directing a sharp "challenging" beam in the direction of the airplane on which a check-up is desired. The aircraft respondor beacon circuit is to contain a double-pulse gate. The gate is tied in with the aircraft distance indicator, so that it varies with distance. The interrogating pulses, beamed from the ground, are spaced automatically as the operator selects the distance range of the target he wishes to challenge. Thus the challenge is narrowed down to distance and direction.

The complete system proposes one airborne transmitter and two receivers. This does not include communications equipment.

MICROWAVE OMNIDIRECTIONAL RADIO RANGE (Sperry). A radio range operating on the same general principles as described for the CAA VHF omnidirectional range, but in the microwave frequencies (2600 Mc), has been developed by the Sperry Gyroscope Company. Like the CAA system, it utilizes the comparison of phase between two audio waves in the airplane to determine the azimuth bearing to the station. In the microwave system, however, the antenna of the ground station is so small that it can be rotated at 1800 rpm, thereby avoiding the use of a capacity goniometer and avoiding antenna phasing problems.

In this system, as in the Sperry microwave instrument landing system, the cw signal to be radiated is generated by a crystal-controlled klystron tube. A synchronous 1800-rpm motor drives a small alternator, the output of which frequency-modulates a 70-kc subcarrier. The subcarrier, as in the CAA system, amplitude-modulates the microwave carrier. The modulated carrier is conducted to an antenna system consisting of a vertical stack of three small loop antennas. The stack gives vertical directivity for maximum horizontal coverage. The modulated carrier radiation produces the reference audio voltage in the airborne receiver. Its relative phase is the same in all azimuthal directions.

The vertical stack of antennas also has a directive pattern in the horizontal plane. This horizontal pattern is essentially sinusoidal. By rotating the pattern with the same motor as used above, another audio voltage is generated in the receiver. This voltave wave, with respect to the former reference voltage wave, has a phase that varies with position around the station. For the full 360° around the station there is a complete cycle or 360° phase variation in the wave. This is called the variable signal. By comparing the phase between this variable and the former reference signal the bearing from the station can be determined.

The phases may be compared in the same manner as in the case of the CAA range, or the phase detector may be used to control the operation of a motor which turns a map in accordance with movement of the airplane around the station.

The proposal for the complete system includes the addition of distance-measuring signals and communications through the single ground station transmitter and antenna.

The greatest advantage in the use of the proposed system is the elimination of the goniometer through the rotation of the antenna system. Reduction in the amplitude of course bends due to siting (reflections from buildings, trees, etc.) are evident.

AEROTRONICS (Raytheon). The Aerotronics system is a proposal based on radar technique and uses PPI scope presentation in the airplane and on the ground. It includes (1) airborne radar, (2) ground beacons (transponders), (3) ground radar, (4) airborne transponders, and (5) distance-measuring equipment. With this equipment, navigation, collision prevention, airways traffic control, approach, and landing are to be taken care of without other aids.

The airborne equipment includes a rotating antenna for radar search, an omnidirectional antenna for communications, and an antenna for distance measuring. The ground radar includes both azimuth and elevation search. This is for traffic surveillance and control. Ground beacons (transpondors) would be used for route navigation.

The distance-measuring equipment proposed operates on the principle of the optical interferometer. It operates a counter at the ground traffic control station where distance to the airplane may be observed with great accuracy.

MULTIPLE TRACK RADAR RANGE (Australian). A multiple track radar range (MTR) has been developed and successfully demonstrated by the Council for Scientific and Industrial Research of Australia at Sydney. This range produces visually indicated (left-right), positively identified flight tracks, based on the time of arrival of pulses from two spaced ground stations. VHF (212 Mc) is employed, which results in line-of-sight distance coverage. The tracks are hyperbolic in shape, but essentially straight beyond 20 miles. The tracks or courses are flyable to within about $\pm 1°$.

The MTR system is fundamentally the same as GEE, a British development used extensively in the invasion of Europe, except for type of presentation. GEE is a hyperbolic system similar to Loran except for its frequency. Two stations only are used in the MTR system, one a master, sending a series of equally spaced pulses (5000 pps in this case) of peak power about 10 kw. The second, a slave station, is equal to the master, but spaced about 8 miles distant and normally inoperative. The slave station receives the pulses from the master and rebroadcasts them with a suitable fixed time delay. For any position around the pair of stations, the pulses are received in a specific time relationship. The contours of equal time differences are hyperbolas passing between the stations and are used as tracks or courses. Non-directional, vertically polarized antennas are used.

In the airborne equipment, the pulse signals are compared in an automatic circuit similar basically to that in the distance-measuring equipment described before. Standard time differences chosen as the tracks are set up on a calibrated switch dial and numbered (track numbers 1 to 30). When a given track is selected, a left-right indicator signals the pilot any deviation from that track.

The master and slave stations are separated in the receiver by permitting the master (only) to transmit double pulses. MTR equipments at adjacent places are identified by a selected difference in repetition rate.

Distance-measuring equipment, of the interrogator-respondor type, is proposed for use with the MTR system.

7. PROPOSED NEW LONG-RANGE NAVIGATION SYSTEMS

CAA LOW-FREQUENCY OMNIDIRECTIONAL RANGE. The low-frequency omnidirectional range operates like the VHF omnidirectional range, i.e., in the comparison of the phase of two audio signals. In early tests of this range, two individual, basic, carrier frequencies were used—one at 172 and one at 194 kc.

The low-frequency omnirange consists of the conventional five-tower Adcock antenna array, one tower at each corner of a square and the fifth tower in the center. The reference signal, 172-kc carrier, modulated with 30 cps, is fed into the center tower only and is, therefore, radiated non-directionally. The other signal (194 kc) is fed to the corner towers through an inductive goniometer which is mechanically rotated at 1800 rpm. The rotation of the goniometer spins the figure-of-eight pattern of the corner towers, producing the variable phase signal in the aircraft receiver. When compared to the phase of the reference signal in a wattmeter circuit, it can be made to indicate azimuth.

Tests are being conducted using a single low frequency and a subcarrier of 1000 cps. The 30-cps reference is applied to this subcarrier by frequency modulation, as in the case of the VHF.

The wattmeter indicating circuits are the same as those used with the CAA VHF omnireceiver.

Night effect represents a serious problem for radio ranges. The lower the frequency employed, the greater is the ground wave signal strength and the more stable will be the operation of the system. Night effect is minimized by curtailing the amount of unwanted vertical radiation. The fact that the signal arrives via ground or sky does not, in itself, introduce an error in the omni system since the indication does not depend upon the length of the path. The development may result in much more freedom from night-effect errors and swinging than in an automatic direction finding system employing a loop and operating on the same frequency.

SONNE (Consul). This is a German aural CW system used against the Allies during the war and capable of great accuracy over long distances and useful in both air and marine service. For its description, refer to Marine Aids, article 10.

LORAN (*lo*ng-*r*ange *a*ir *n*avigation). Excellent results are being obtained in the extensive use of Loran in transoceanic flights. For a description of this system, refer to Marine Aids, article 10.

NAVAGLOBE (Federal). This is a CW system using very low frequencies for long-range navigation over oceans and continents. The receiver proposed has a narrow-band, noise-rejecting feature and ADF facilities. The ground station consists of three antennas spaced in the corners of an equilateral triangle, and a transmitter which is connected in succession to the three antenna pairs each second. Three dumbbell-shaped patterns, displaced in bearing by 120° from each other, are radiated successively. Relative amplitudes of the three successive signals received during each cycle are measured automatically by a ratiometer, permitting direct visual indication of bearing at all azimuths from the station. For other directions between courses, measurement of ratio in a ratiometer in the receiver permits determination of relative bearing with respect to the antenna array.

8. MISCELLANEOUS RADIO AIDS

VERTICAL SEPARATION INDICATOR. The Stratoscope is an instrument conceived to provide visual indication in an airplane and on a ground monitor of the vertical separation between airplanes (or other obstacles) within a minimum service area of 10 miles distance and 1000 ft elevation.

Basically, the Stratoscope operates by converting aircraft height into frequency and then utilizing panoramic reception to display received signals along a CRT time base calibrated in relative vertical height. The airplane equipment includes a transmitter and a receiver-indicator unit. The ground monitor needs only a receiver-indicator unit.

Conversion of height into frequency is accomplished by means of a precision aneroid cell which operates a variable tuning condenser in a coaxial line oscillator. The oscillator frequency changes 5 Mc in 10,000 ft. This change is accomplished in the frequency band of 148–154 Mc. The height of the transmitter in relative pressure altitude may therefore be determined by measurement of transmitter frequency.

Interference from the plane's own transmitter is avoided by sharing time between the transmitter and receiver at an alternation rate of 30 cps. By coding the transmission, some identity of the airplane can be provided in the receiving indicators.

In tests of this equipment an accuracy of 200 ft was observed. The required 10- or 20-mile distance range can be covered with about 5 watts (transmitter power output).

ABSOLUTE ALTIMETERS. Absolute altimeters have little utility in navigation over land because of the rapid variation in indication in rough territory, but they do have utility in navigation over water, which is flat, as a means of maintaining constant height above the surface. Variations then in the pressure altimeter indicate flight toward or away from atmospheric-pressure areas. Flights directed laterally so as to give constant-pressure altimeter reading (for constant absolute altitude) avoid storm areas over the oceans.

Many absolute altimeters have been invented over a period of twenty years, including sonic, capacity, frequency modulation, and radio-pulse types. A reference is given for detailed information as space permits only summary remarks here.

One absolute altimeter, called a "terrain clearance indicator," was developed for commercial use just before the war. It operates on the FM principle at about 432 Mc. The transmitter continuously radiates its energy downward from a doublet antenna. The frequency is varied from 410 to 445 Mc sixty times per second. The receiver takes in the energy reflected back from the ground and also some of the original transmitted energy. The receiver output, which may be considered a beat between the two signals, has a frequency depending upon and increasing with altitude.

The FM type absolute altimeter is capable of operating down essentially to zero altitude. It has greatest accuracy at lowest altitude but is subject to possible serious errors from adjustment and noise which prevent its immediate use on instrument landing.

The radiopulse altimeter developed during the war for high-altitude bombing uses a cathode-ray tube as a means of displaying time between transmitted pulse and echo. The display is calibrated in feet of altitude and arranged circularly around the tube face. Essentially it is a radar distance indicator with its antenna fixed in a position under the airplane. The pulse type altimeter is extremely accurate at high altitude but, at present, useless at altitudes below several hundred feet. Future development may bring about reduction in this minimum altitude.

REFERENCES

GOVERNMENT—CAA

The U. S. Dept. of Commerce—How It Serves You on Land and Sea and in the Air. U. S. Government Printing Office (January 1946).

The FCC Allocation Plan, *Electronics*, March 1945, p. 92.

Advancing Air Navigation. U. S. Government Printing Office (1946).

FOREIGN

Gravis, Airline Navigation Was Different in Europe, *Aviation Mag.*, November 1940, p. 40.

Hodgson, Aircraft Radio Equipment for Use on European Air Lines, *Proc. I.R.E.*, September 1935, p. 979.

RADIO RANGES—LF

Diamond, On the Solution of Night Effects and the Radio Range Beacon System, *Proc. I.R.E.*, June 1933, p. 808.

Jackson-Stuart, Simultaneous Radio Range and Telephone Transmission, *Proc. I.R.E.*, March 1937, p. 314.

Army-Navy Precipitation Static Project, Parts I to VI, *Proc. I.R.E.*, April 1946, p. 156P; May 1946, p. 234P.

RADIO RANGES—VHF

An Ultra-high-frequency Radio Range with Sector Identification and Simultaneous Voice, *Electrical Comm.*, June 1946, and *Proc. I.R.E.*, January 1946, p. 9W.

Development of the VHF Radio Range, Part III, U. S. Dept. of Comm. CAA T. D. Report 49.

RADIO MARKERS

Jackson, Metz, McKeel, Test of First Manufactured Fan Marker, U. S. Dept. of Comm. CAA T. D. Report 15 (July 1938).

Hromada, Development of New Station Location or Z Marker, *Proc. I.R.E.*, August 1944, p. 63.

LaPort, Radiating System for 75 Mc Cone of Silence Marker, *Proc. I.R.E.*, January 1942, p. 26.

AUTOMATIC DIRECTION FINDERS

Roberts, H. W., *Aircraft Direction Finders*, Chapter 8, p. 190. Morrow, New York.

COMMUNICATIONS

Ellihorn, Anti-noise Characteristics of Differential Mics, *Proc. I.R.E.*, February 1946, p. 84P.
Shawn, Application of the Throat Mic, *Communications*, January 1943, p. 11.
Bennett et al., The Design of Broad Band Aircraft Antenna Systems, *Proc. I.R.E.*, October 1945, p. 671.

AIR TRAFFIC CONTROL AND WEATHER

Air Traffic Rules, CAA Manual 60. U. S. Government Printing Office (October 1945).
Gilbert, Future Air Traffic Control, *Flying Mag.*, April 1946, p. 60.
Diamond, Recent Applications of Radio to Remote Indication of Meteorological Elements, *Elec. Eng.*, April 1941, p. 163.
Hauck, Radiosonde Telemetering Systems, *Electronics*, May 1946, p. 120.
Is the Air Full, *Harper's* (July 1946).

INSTRUMENT LANDING

Instrument Landing of Aircraft, *Elec. Eng.*, December 1940, p. 495.
Metz, The CAA–RTCA Instrument Landing System, U. S. Dept. of Comm. CAA T. D. Reports 35 and 36 (October 1943).
Caporale, The CAA Instrument Landing System, *Electronics*, February 1945, p. 116; March 1945, p. 128.
Metz, Radio Glide Path for Aircraft, *Radio News*, November 1945, p. 8.
Montgomery, A VHF Aircraft Antenna for Reception of 109-Mc Localizer Signals, *Proc. I.R.E.*, November 1945, p. 767.

OMNIDIRECTIONAL RADIO RANGE

Luck, Omnidirectional Radio-Range System, *R.C.A. Rev.*, March 1946, p. 94.
Stuart, The Omni-Directional Range, *Aero Digest* (June 15, 1945).

RADAR

The Radar Equation, *Electronics*, April 1945, p. 92.
Schneider, Radar, *Proc. I.R.E.*, August 1946, p. 528.

NEW LANDING SYSTEMS

Ground Controlled Approach for Commercial Aviation, *Electronics*, May 1946, p. 160.
Spicer, The GCA Landing Systems, *Bendix Radio Engineer*, January 1946, p. 17.
Folland, The Use of Microwave for Instrument Landing, *Radio*, April 1946, p. 23.
Aerial Navigation and Traffic Control with Navaglide, etc., *Electrical Communication*, June 1946, p. 113.

ROUTE NAVIGATION

Herbst et al., The Teleran Proposal, *Electronics*, February 1946, p. 124.
Aerial Navigation and Traffic Control with Navar and Navaglobe, *Electrical Communication*, June 1946, p. 113.
Pierce, Introduction to Loran, *Proc. I.R.E.*, May 1946, p. 216.
Sandretto, Absolute Altimeters, *Proc. I.R.E.*, May 1944, p. 167.
Frequency, Power and Modulation for Long-range Radio Navigation System, *Electrical Communication*, June 1946, p. 144.

RADIO AIDS TO MARINE NAVIGATION

By M. K. Goldstein

Table 1 shows an extensive summary and classification of established, recently introduced, contemplated, and proposed electronic aids to marine navigation. It will be noticed that the systems are classified in groups dealing with radio types, sonic and supersonic types, and red and infrared types. Also the various systems are analyzed for essential characteristics falling into the following grouped categories: Type of system (including basic principle and position information supplied); performance of system (including range, accuracy, ambiguities, operating frequency, and band width; system requirements (including required ship and shore equipments); and the system usability (including engineering status and reliability factors). Details concerning each of the specific systems classified are discussed in the following articles in this section.

Table 1. Summary and Classification of Established, Recently

System	Type: Principle Employed	Type: Basic Information Supplied	Performance: Maximum [1] Range	Performance: Accuracy [2]	Performance: Ambiguities	Performance: Frequency (Mc/s)
RADIO						
1. *Beacon:*						
a. Radiobeacon (shore)	Non-directional transmission and directional reception	Azimuth	Short, med.	<2°	None	0.285–0.315 [3]
b. Radiobeacon (ocean sta.) †	Non-directional transmission and directional reception	Azimuth	Medium	<2°	None	0.30–0.55
c. Radar marker (Ramark)	Non-directional transmission and directional reception	Azimuth	Optical	Devel'pm't'l	None	3, 5, & 10 cm.
d. Respondor (Racon)	Non-directional transmission and directional reception	R & az. fix	Optical	±2%, ±3°	None	3, 5, & 10 cm.
e. Reflector (Racon).	Non-directional transmission and directional reception	R & az. fix	Optical	±2%, ±2°	None	3, 5, & 10 cm.
f. Rotating beacon (Consol.)	Dir. trans. & non-dir. reception	Azimuth	Medium	1.7°–2.3°	Multiple [4]	0.20–0.5
2. *Direction finding—ship* (Loop):						
a. MF—Aural null (Standard) †	Directional reception of any transmission	Azimuth	Normal communication range	<2°	None	0.3–3.0
b. MF—Null seeking (ADF) (SCR-269) †	Directional reception of any transmission	Azimuth	Normal communication range	<2°	None	0.2–1.75
c. MF-Instantaneous (DAK) †	Directional reception of any transmission	Azimuth	Normal communication range	<2°	None	0.25–1.5
d. MF—Position plot (Bendix)	Directional reception of any transmission	Az. fix	Normal communication range	<2°	None	0.1–1.6
e. HF—Instantaneous (DAQ, DAU) †	Directional reception of any transmission	Azimuth	Normal communication range	5°–10°	None	1.5–21
3. *Direction finding—shore:* †						
a. MF—Adcock (DAH)	Directional reception of any transmission	Azimuth	Normal communication range	1°–2°	None	0.25–1.5
b. HF—Adcock (DAJ,SCR-291)	Directional reception of any transmission	Azimuth	Normal communication range	3°–4°	None	1.5–30
c. HF—spaced loop (DAB)	Directional reception of any transmission	Azimuth	Normal communication range	3°–4°	None	2.0–18
d. VHF—Adcock (DBF)	Directional reception of any transmission	Azimuth	Optical	3°–5°	None	100–160
e. UHF—Reflector (DBM)	Directional reception of any transmission	Azimuth	Optical	3°–5°	None	90–5000
4. *Radar:*						
a. Ship......... ..	Echo ranging	R & az. fix	Optical	±2%, ±2°	None	3, 5, & 10 cm.
b. Shore......	Echo ranging	Traffic	Optical	±2%, ±2°	None	3, 5, & 10 cm.
5. *Propagation time difference:*						
a. Loran (standard).	PTD [6]—synch. pulsed transmissions	Hyp L of P [7]	Long	0.5%	None [8]	1.7–2.0
b. Loran (SS) [10].....	PTD [6]—synch. pulsed transmissions	Hyp L of P	Long	0.5%–1.0%	None [8, 18]	1.7–2.0
c. Loran (LF) [11]....	PTD [6]—synch. pulsed transmissions	Hyp L of P	Long	<1.0%	None [8]	0.18
d. Gee.............	PTD [6]—synch. pulsed transmissions	Hyp fix	Optical	<0.8%	None [8]	20–85
e. Decca.....	PPD [13]—Synch. CW transmissions	Hyp fix	Medium	±0.02% [14]	Multiple [15]	0.01–0.2
f. POPI...... . .	PPD [13]—Synch. CW transmissions	Hyp L of P	Long	Devel'pm't'l	None	0.75
6. *Interrogator-respondor:*						
a. Shoran...........	Dual beacon ranging	Range fix	Optical	±50 ft	Few [16]	210–320
b. Lanac—ship......	Single beacon ranging	R & az. fix	Optical	±2%, ±2°	None	1000
c. Lanac—shore	Single beacon ranging	Traffic	Optical	±2%, ±2°	None	1000
7. *Composite data relay:*						
a. Teleran...........	Instant. data relay	Gen. nav.	Optical	....	....	Telv. Ch'n'l
b. Facsimile.	Recording data relay	Gen. nav.	Medium	....	....	Facs'm'le Ch'n'l

Introduced, Contemplated, and Proposed Aids to Marine Navigation

Performance	Requirements		Appraisal Factors *	
Signal Band Width	Minimum Equipment for Basic Information Supplied: Ship	Minimum Equipment for Basic Information Supplied: Shore	Engineering Status	General Remarks
≥CW	DF receiver	Coded beacon	Proven	Inaccurate and unreliable beyond predominant ground wave range.
≥CW	DF receiver	Coded beacon	Proven	Combination radiobeacon, rescue ship, weather station.
3–5 Mc	Special radar	Beacon	Under development	Identified by reference to natural target echoes.
3–5 Mc	Special radar	Respondor beacon	Proven	Identification obtained by coded transmission.
3–5 Mc	Radar system	Reflector	Proven	Special target for increased echo reflection.
≤1 kc	Receiver	1 station	Under trials [5]	Minimum range—25 mi; accuracy deteriorates at night beyond predominant ground wave range.
≥CW	Standard DF	CW or other transmitter	Proven	2a to 2e inclusive—unreliable beyond predominant ground wave range due to susceptibility to polarization error. Note to 2d: Fix obtained by simultaneous cross-bearing plot on map.
≥CW	Special DF		Proven	
≥CW	Special DF		Proven	
≥CW	2 ADF'S & computer		Under development	
≥CW	Special DF		Proven	
≥CW	MF transmitter	Special DF	Proven	3a to 3e inclusive—designated models have low susceptibility to polarization errors beyond ground wave ranges. Indicated accuracies generally realized up to maximum sky wave range.
≥CW	HF transmitter	Special DF	Proven	
≥CW	HF transmitter	Special DF	Proven	
≥CW	VHF transmitter	Special DF	Proven	
≥CW	UHF transmitter	Special DF	Proven	
3–5 Mc	Radar system	None	Proven	Valuable anti-collision and short-range naval device.
3–5 Mc	Comm. receiver	Radar, comm. transmitter	Under trials	Suitable for traffic control in congested areas.
50–70 kc [9]	Special receiver and indicator	2 stations	Proven	Day range, 750 mi; night range, 1400 mi (over sea water).
50–70 kc [9]		2 stations	Under trials	Night use only; utilizes base line of approx. 100 mi.
10 kc		2 stations	Under trials	Greater ground wave range and more stable propagation over 5a.
1 Mc [12]		3 stations	Proven [5]	Simultaneous fix from 2 hyperbolic lines of position.
3 CW freq's.	Special receiver	3 stations	Under trials [20]	Limited as 1(a) above owing to ionospheric phase shifts.
≤1 kc	Rcvr. & Spc'l ind.	1 station	Under development [5]	Same as 5e; awaits satisfactory phase measurement between sequentially received signals.
3-5 Mc	Special trans. & receiver	2 respondor beacons	Proven [17]	Bombing and mapping aid; may have marine use.
3–5 Mc		Respondor beacon	Under trials	Optional operation: radar beacon or radar system.
3–5 Mc	Respondor beacon, comm. receiver	Spc'l. trans. & rcvr., comm. trans.	Under trials	Traffic control utilizing shipboard transponders.
3–5 Mc	Television receiver	Television transmitter	Under development	Air traffic control system; also feasible for marine use.
1–10 kc	Facsimile receiver	Facsimile transmitter	Proven	Permanent recording of any general information, data, etc.

Table 1. Summary and Classification of Established, Recently Introduced,

System	Type		Performance			
	Principle Employed	Basic Information Supplied	Maximum [1] Range	Accuracy [2]	Ambiguities	Frequency (Kc/s)
SONIC AND SUPERSONIC						
8. *Underwater beacon*	Non-dir. trans. & dir. reception	Azimuth	V short	<1°	None	0.8–30
9. *Direction finding*	Dir. reception of any transmission	Azimuth	V short	±1°	None	0.8–30
10. *Sonar:*						
a. General echo ranging	Gen. echo ranging	R & az. fix	V short	<1%, <1°	None	15–30
b. Echo sounding	Vert. echo ranging	Depth	V short	<1%	None	0.8–30
11. *Sofar* †	PTD [6]—Explosion & multiple timed reception	Position fix	V long	<5 mi	None	0.05–1.0
RED AND INFRARED						Microns [19]
12. *Beacon*	Non-dir. trans. & dir. reception	Azimuth	V short	<1°	None	Far Infrared (0.7–13)
13. *Direction finding*	Dir. reception of any transmission	Azimuth	V short	<1°	None	Far Infrared (0.7–13)
14. *Redar*	Optical echo ranging	R & az fix	U short	Proposed	None	Near Infrared (0.3–1.5)

* Based on data available up to late 1949.
† Used entirely or in part for distress service.

[1] Ultra-short	0–1 nautical miles
Very short	1–10 nautical miles
Short	10–150 nautical miles
Medium	150–500 nautical miles
Long	500–1500 nautical miles
Very long	over 1500 nautical miles

[2] Representative accuracies obtained in practice.
[3] Frequencies in current U. S. use.
[4] Resolved by dead reckoning or DF.
[5] British only.
[6] Propagation time difference.
[7] Hyperbolic line of position.
[8] Two easily resolved ambiguities can exist in fix obtained.
[9] Provisions are made for 16 different channels on the same carrier frequency by utilizing different pulse repetition rates.

9. ESTABLISHED NAVIGATIONAL AIDS

RADIOBEACON SYSTEMS (MF) (See references 2 and 3.) The non-directional medium-frequency (MF) radiobeacon system, maintained by the Coast Guard in the United States and by similar organizations in foreign countries throughout the world, is the most extensively used radio navigational aid today. The system consists of fixed radiobeacon stations located at lighthouses, lightships, and other points, which transmit distinctive identifying coded tone signals, enabling navigators at sea to take bearings on them by means of medium-frequency loop direction finders. Bearings obtained from two or more such beacon stations, in conjunction with charts of the geographical locations of the radiobeacons, uniquely establish the ship's position. A single radio direction-finding (DF) bearing crossed with a line of position of a heavenly body, two bearings on the same station and the distance run between bearings, or a bearing and a synchronized air or submarine fog signal also suffice to determine the position of the vessel. In the last case, blasts from the sound signal are synchronized with the radiobeacon signals and the difference in time of reception of these two signals can be converted into an approximate station distance. Owing to the many factors which enter into the transmission and reception of radio signals, a ship cannot estimate its distance accurately from a radiobeacon either by the strength of the signal received or by the time at which the signals were first heard. However, a line of direction obtained from a single radio bearing enables a ship to proceed toward the radio station by the shortest course. This is especially applicable to a rescue ship, enabling it to head directly toward the ship in distress and thereby arrive in the minimum of time. Such radio bearings are usually accurate to within 2° or less, depending upon the equipment and the operator's skill. DF calibration transmissions are available from radiobeacons upon request, continuous signals being transmitted while the ship's direction finder is being calibrated. The simplicity and reliability of the radiobeacon indicate that such systems will continue to exist as a navigational aid for many years.

The present MF radiobeacon system in the United States (see Fig. 1 for East Coast and Gulf) consists of approximately 200 radiobeacons, each radiobeacon operating on one of the even frequencies in the 285–315 kc/s band. The European system frequency

Contemplated, and Proposed Aids to Marine Navigation—*Continued*

Performance	Requirements		Appraisal Factors *	
Signal Band Width	Minimum Equipment for Basic Information Supplied		Engineering Status	General Remarks
	Ship	Shore		
≥CW	Sonic DF	Sonic beacon	Proven	Usefulness limited by certain underwater phenomena; see 10(a).
≥CW	Sonic DF	Any sonic transmission	Proven	Bearing on any received underwater sound transmission; see 8.
1–4 kc	Special transmitter and receiver	None	Proven	Accuracy subject to underwater refraction and propagation.
1–4 kc	Special transmitter and receiver	None	Proven	Depth finding by reflected sound transmissions.
<1000 cyc	Special explosive charge	Special receiver stations	Under trials	Proposed distress aid using natural underwater sound channel.
Microns				
Approx. 13	Infrared DF	Infrared beacon	Proven	Same range as visual light under fog conditions.
Approx. 13	Infrared DF	Any infrared transmission	Proven	Same range as visual light under fog conditions.
1.2	Special transmission and receiver	None	Proposed	Same range as visual light under fog conditions.

10 Intended to increase nighttime accuracy over Standard Loran; however, comparable accuracy has not yet been practically achieved.
11 Intended for increased daytime range over Standard Loran; however, comparable accuracy has not yet been fully realized.
12 If necessary, different pulse repetition rates will permit multistation operation on same carrier frequency.
13 Propagation phase difference.
14 For medium range.
15 Resolution awaits satisfactory system of "lane" identification.
16 Solved by general knowledge of position.
17 Air only.
18 Four-station operation also provided to give increased accuracy and to remove ambiguities.
19 The thousandth part of 1 mm.
20 A British Isle Decca facility is in 24-hr operation with performance results indicating a proven system.

band is 290–320 kc/s. Distinctive signals, by which the different stations are positively identified, are obtained by keying a tone-modulated carrier to give simple dot-dash combinations at the rate of thirty characters per minute. Because of the great number of stations and the restricted frequency band, it has been found necessary to share time between stations on the same radio frequency by transmitting signals from any one radiobeacon for 1 minute followed by 2 minutes of silence. This allows two adjacent stations to transmit similar signals on the same frequency without causing interference simply by timing the transmissions to occur during successive minutes. Two-tone modulation is used on some stations in congested traffic areas to add distinctiveness to the identifying signal. During clear weather the radiobeacons follow such cycles of transmission for one or two 10-minute periods of the hour. All six 10-minute periods are utilized during conditions of fog. The beacons are carefully timed and remotely monitored.

Radiobeacons are grouped into four classifications, according to their power output and maximum effective ranges:

Class A	750 watts—200 miles
Class B	150 watts—100 miles
Class C	25 watts— 20 miles
Class D	5 watts— 10 miles

A few special high-power beacons having ranges of 400 miles have also been put into operation. Class A and B stations require identical equipment with the exception of the 750-watt power amplifier used with the A stations. The transmitter, which is coupled directly to the antenna for the B station, is used as an exciter for the power amplifier of the A station. Both types of radiobeacons often use 125-ft insulated self-supporting towers as non-directional antennas, with provision for location up to several hundred feet from the transmitter house when local conditions make such an installation preferable. Sometimes an insulated guyed antenna, approximately 40 ft high, is placed on top of a light tower above the lantern. Lightships use vertical or symmetrical T antennas. Inverted L-type antennas are not used because of the undesired directivity and the horizontally polarized field components. The frequency of the A and B stations can be quickly changed to any one of four pre-set frequencies which are usually crystal-controlled. Class C radiobeacons are used for short-range navigation. The equipment is very similar to that of a class A station except for the much lower power and the reduction in size of equipment. Owing

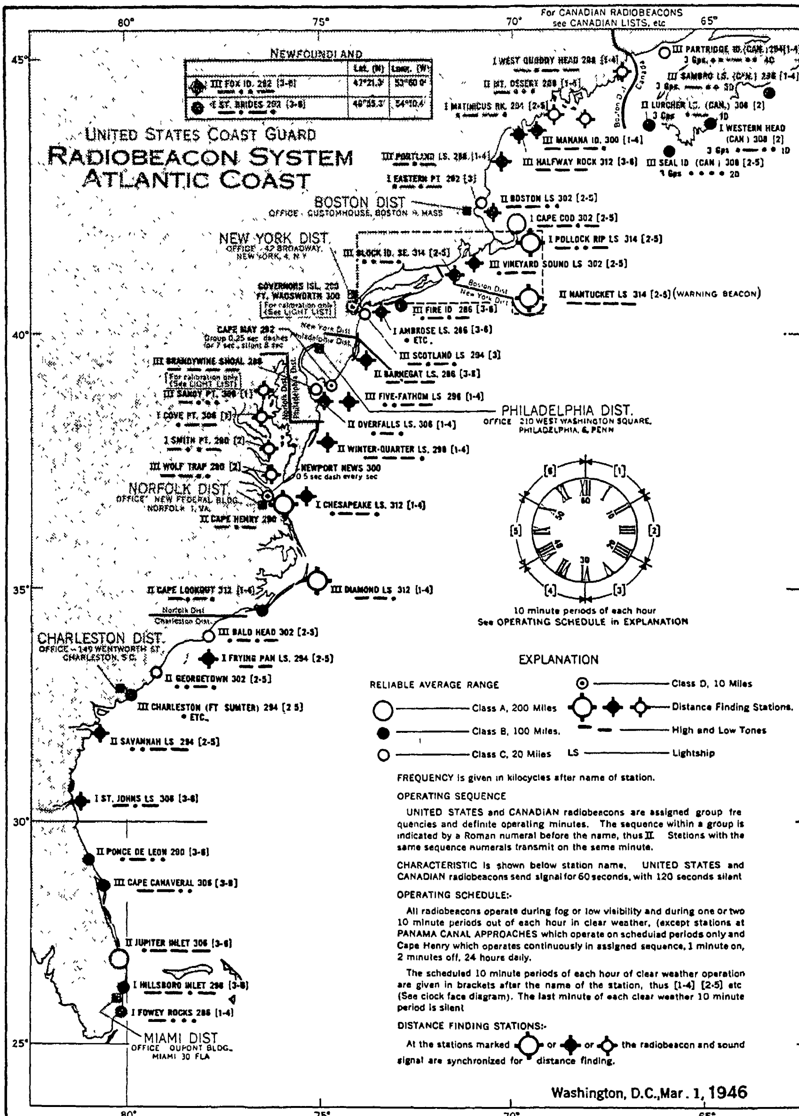


FIG. 1. Typical Radiobeacon Navigational Chart, Published by U. S. Coast Guard

to the necessity of making the signals from these short-range stations especially distinctive in areas of considerable marine traffic, the modulating audio frequency is variable in four steps. The timer cams can be set to give a single tone or any combination of tones. Class D marker radiobeacons are small automatic battery-operated transmitters located on pierheads, buoys, ends of jetties, etc., which serve as local markers to indicate channel entrances, turning points, etc., where careful approach is required. They are not synchronized with other radiobeacons but operate continuously in all types of weather, sending out a characteristic of several dashes during each 30-sec period. The transmitter frequency is crystal-controlled and has a modulation frequency of 1000 cycles per second. A 15-ft welded Monel tripod mast is normally used as the antenna. Special warning transmitters are located at some radiobeacons; each transmitter operates on the same frequency as the beacon and gives a distinctive signal to warn a navigator who is homing into the station that he is approaching dangerously close. The usable range of these warning beacons is, therefore, intentionally reduced to provide warning over only a small but sufficient area surrounding the long-range beacon.

At present the characteristics of some of these stations are being changed to provide satisfactory operation with automatic direction finders. Instead of both the carrier and modulation being interrupted during the keying process, only the modulation will be interrupted. In addition, many of the stations will transmit throughout the hour under all weather conditions rather than just during one or two 10-minute periods.

The well-established aerobeacon system used for air navigation is also used for marine purposes and is discussed in the first part of this section.

DIRECTION-FINDING SYSTEMS. (See references 4, 5, and 6.) The use of shipboard medium-frequency radio direction finding has continued to grow as a navigation aid since the early days of radio broadcasting. By use of the shipboard direction finder, practically every transmitting station within the frequency band and range of the radio direction finder is a potential point for a navigational reference line of position. Shipboard direction finding also plays an important role in directly guiding vessels to other vessels in distress. As an additional aid to navigation in time of distress, medium-frequency direction-finding stations located along the coasts and on the Great Lakes are available to mariners who transmit signals on the international distress frequency of 500 kc/s. Various other frequencies are available for distress and emergency purposes, depending on the range from land or the location of the vessel. For example, 8280 kc/s is used for United States long-range contact and 2182 kc/s is used on the Great Lakes.

While the most widely used shipboard direction finder is of the loop aural-null type, other types are available (see Section 6, article 32).

Environmental Effects. (See reference 5.) The term "environmental effects" is used to describe all local physical conditions which cause DF bearing errors aboard ship and at shore stations. Such conditions normally involve metallic structures near the DF antenna and their radiation fields, induction fields, or shielding effects. For frequencies below 1000 kc/s, shipboard environmental effects can be sufficiently well controlled or compensated (further details are given by C. T. Solt, *Proc. I.R.E.*, Vol. 20, p. 228, February 1932) so that direction finders in this band give highly satisfactory performance on ground-wave transmissions; i.e., the resultant calibration curve is fairly symmetrical and does not exceed a few degrees maximum deviation. Arriving signals of certain frequencies cause partial or complete resonances in, and consequent reradiation and induction fields from, nearby metallic structures. These fields generally impart 90° phase components (called quadrature effects) to the arriving wave fronts, altering their apparent arrival direction by amounts depending on the positions and sizes of the resonating structures. These changes in direction result in deviations of the observed bearing from the true bearing. Deviation, then, is defined as the quantity that "must be added algebraically to the observed bearing in order to make it equal the correct bearing"; the deviation, however, does not affect the resolving power (angular sensitivity) of the DF. The quadrature effect, on the other hand, causes an elliptical polarization condition of the E or H fields, or both, which generally results in failure of the antenna to find a sharp null coupling position to these fields, thus adversely affecting the resolution of the DF. Figure 2(a) shows the maximum deviation that may be expected due to induced currents at frequencies at which the structure may be resonant (worst case). For frequencies off resonance and $f \lessgtr 0.5$ the shapes of the curves of Fig. 2(a) are maintained, but the magnitude of the deviation is reduced approximately as follows:

$$\frac{D_{\text{res.}}}{D} = \frac{e/r}{e/(r + jx)} = 1 + j\left[1 - \left(\frac{f_0}{f}\right)^2\right]Q \tag{1}$$

or

$$D = \frac{D_{\text{res.}}}{1 + jQ[1 - (f_0/f)^2]} \tag{2}$$

where e is the voltage induced in the structure at frequency, f; r, x, L, C, and Q are the electrical constants of the structures; $D_{\text{res.}}$ is the maximum deviation at the resonant frequency f_0; and D is the maximum deviation at any frequency. The resonant frequency, f_0, of structures may be determined from Fig. 2(b), to which some corrections should be

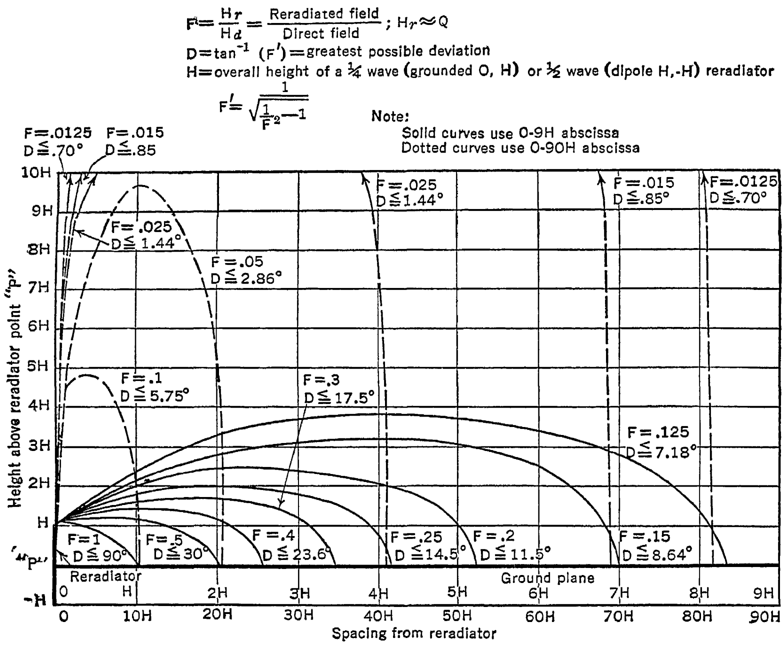


Fig. 2(a). Bearing Deviations Due to Presence of a Reradiator

Notes. 1. D (Worst Possible Deviation) is directly dependent upon the lateral and vertical spacings from the reradiator. 2. A resonant reradiator of $Q = 20$ has been assumed. 3. For other Q's and non-resonant conditions, see text. 4. Curves show reradiator spacings to maintain constant D or F.

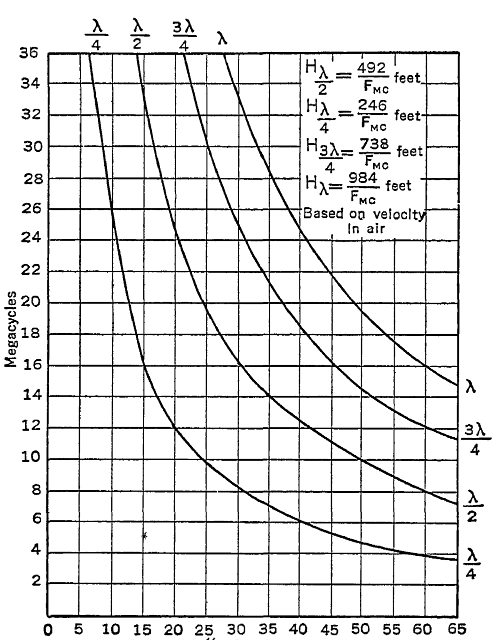
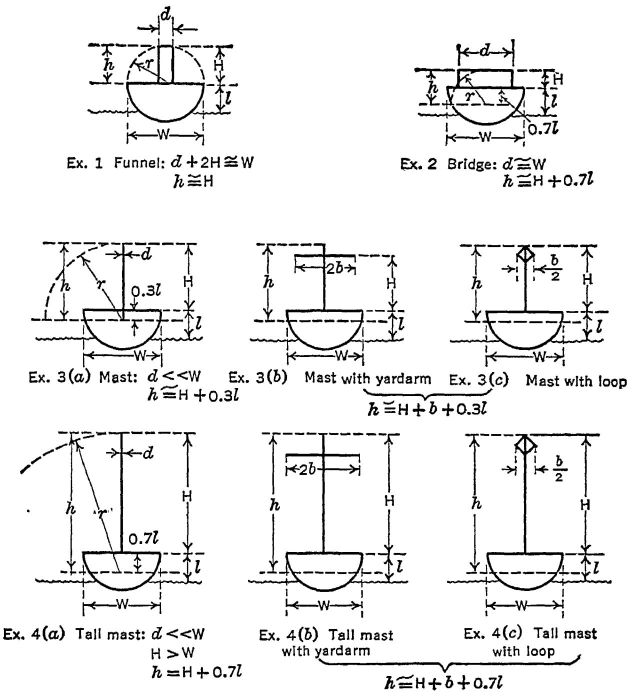


Fig. 2(b). Fundamental and Harmonic Resonant Frequencies of Conductors of Physical Heights H

added for end effects and top loading as shown in Fig. 2(*c*). The range of *Q*'s is approximately as follows:

Wide-width reradiators (superstructures)	1.5– 2
Medium-width reradiators (funnels)	2 – 3
Narrow-width reradiators (masts)	5 –10
Very thin reradiators (guy wires, antennas)	10 –20

It can be seen from these results that DF antenna locations on top of the tallest structures give the least reradiation error effect from those structures and, generally, maximum clearance (least error) from the other, surrounding structures. Breaking up structures



Note: *r* shows region of strong reradiated fields

FIG. 2(*c*). Top Loading and End Effect Corrections for Shipboard Reradiator Heights *H*

(e.g., guy wires, rails) with insulators, to reduce the induced currents, further reduces the environmental effects. Partial compensation of environmental effects can often be obtained by judiciously introducing compensator loops or structures, such that an approximately equal and opposite (compensating) environmental effect is obtained. Successful means have been devised for altering the frequency characteristics of structures surrounding a DF antenna in order to reduce and control environmental effects. (See reference 5.)

Quadrature effects are generally reduced by minimizing the environmental effects as stated above. However, in practical direction finders, some polarization effects and residual environmental effects often leave high values of quadrature effect which must be further minimized in order to realize satisfactory DF bearing resolution. Quadrature balance or compensation has long been used for this purpose. It makes use of a non-directional antenna for obtaining a voltage from the arriving wave which, with a properly adjusted phase and amplitude, can be made to cancel out completely the undesired quadrature voltage derived from the directional collector. In practice, the phase shift, if any, required for the above balancing or compensating voltage is a constant value over the frequency band. As a result, most direction finders operating below 2000 kc/s incorporate this balance circuit with only a single associated panel control for positive and negative magnitude of compensation adjustment. The non-directional antenna employed for balancer purposes is actually the same one generally furnished in practically all direction finders for sense determination, i.e., for resolving the 180° ambiguity in the direction of a received signal. When used for that purpose, the phase and magnitude of the sense antenna voltage is adjusted (generally by a fixed circuit) to match (as closely as possible) the phase and magnitude of the directional antenna's maximum coupling output. In this manner the cosine (double null) coupling law is converted to an approximate cardioid (single null) coupling law which possesses the desired single non-ambiguous null. Because the cardioid law yields a poorer bearing resolution, i.e., differential amplitude to angle ratio, in the region of its null as compared to the cosine law null region, the cardioid is rarely used for other than sense determination.

Receiver-introduced modulation for quadrature-effect suppression (see reference 6), which has recently been developed, possesses several advantages over the balancer type. It operates on the principle of phase discrimination for the undesired quadrature component. Phase discrimination is accomplished by introducing the voltage e_1, derived from the non-directional antenna, into the DF as a reference voltage of proper amplitude and phase with respect to the directional voltage, e_2. The directional voltage e_2 is then passed through a mechanical or electronic reversing switch just before it is added to e_1. The resultant e_3 has a maximum $+e_3$ and a minimum $-e_3$ value depending upon the position of the reversing switch as shown in Fig. 3(*a*). It will be noted in Fig. 3(*b*) that the undesired or quadrature directional antenna voltage component e_2', after passing through the common reversing switch, is also vectorially added to e_1. The resultant e_3' thus also has values $+e_3'$ and $-e_3'$ depending upon the position of the reversing switch. However, it will be noted that the desired directional voltage $\pm e_2$ causes a maximum differential effect on e_3, whereas the undesired or quadrature directional voltage $\pm e_2'$ causes a minimum (zero) differential effect on e_3'. Moreover, for small values of quadrature, the magnitude of $+e_3'$ differs negligibly from e_1. Thus, if e_2' is 10 per cent of e_1, $\pm e_3 = \sqrt{(1)^2 + (0.1)^2} = 1.005$, or the presence of 10 per cent e_2' causes only 0.5 per cent increase in e_3 over e_1. By utilizing suitable synchronous commutators, or sweep circuits with a cathode-ray tube, the maximum and minimum resultant voltages $\pm e_3$ can be compared and the desired sharp null position of the collector can be obtained just as readily with considerable quadrature voltage present as when none is present, since the entire quadrature suppression is continuous, automatic, and independent of the operator. It should be noted that positive sense determination is continuously present in this type of system since the non-directional reference voltage, e_1, is automatically a sense voltage, and sense is obtained merely by correlating the instantaneous position of the reversing switch with an increasing (or decreasing) resultant output. It can be shown that the receiver-introduced-modulation principle markedly improves the signal-to-noise ratio (or the bearing sensitivity) by integrating out the random (non-synchronous) noise effects. It can be effectively demonstrated that satisfactory DF bearings can be obtained on broadcast program modulated signals notwithstanding noise levels that obliterate the program intelligence. The cathode-ray tube comparator for "on bearing" indication in some U. S. Navy direction finders is especially desirable for obtaining indication on ICW transmissions. When CW or MCW transmissions are employed the entire DF bearing may be taken automatically by employing a suitable controlled servo system to orient the directional collector for minimum difference in $+e_3$ and $-e_3$. This identical principle is employed in the automatic direction finder (ADF) described in article 3 under Aids to Air Navigation, and as a second mode of operation in the Navy DBD MF/DF system, the latter being especially designed for simplified marine use. Both the balancing and the modulator suppression of quadrature effects may, under certain conditions, introduce a deviation effect. This may occur when a large quadrature (undesirable) voltage, e_1', is induced in the non-directional antenna. This condition, and its amount, can be quickly identified by a shift of the reciprocal bearing (i.e., observed bearing plus 180°) from its normal 180° position. Since this effect is almost identical to, and originates from, the same mechanisms responsible for the normal Deviation, it is as stable as the latter, and can be absorbed by the overall DF calibration when made.

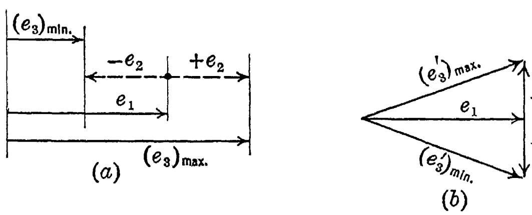


FIG. 3. Vector Relations for Quadrature-effect Suppression by Modulation Method

Calibration Curves. A conventional DF calibration curve is shown in Fig. 4(*a*) which is generally taken for a few of the most received frequencies. The same data can be presented in the automatic interpolation (see reference 5) form shown in Fig. 4(*b*). The latter form permits more direct plotting and utilization of the calibration data with the great

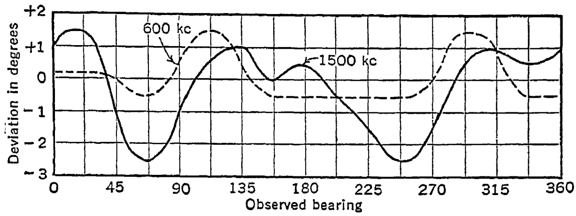


FIG. 4(*a*). Conventional Deviation Chart for Direction-finding Calibration

advantage that the corrected bearing of other than calibrated frequencies may be ascertained just as conveniently as that of the calibrated frequencies.

Polarization Effects. Radio waves propagated with their electric field vertical have become known as *vertically polarized* waves; those having their electric field horizontal

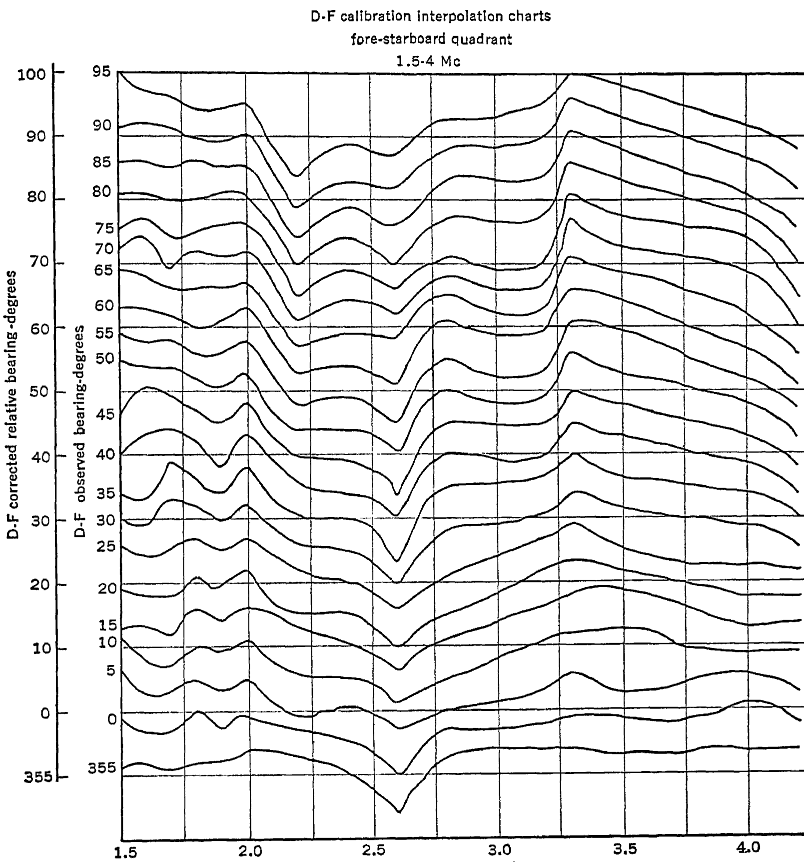


FIG. 4(*b*). Automatic Interpolation Chart for Direction-finding Calibration

have become known as *horizontally polarized* waves; those having their electric field inclined are a combination of both types. Experience shows that most direction finders are susceptible to appreciable unwanted polarized energy pickup, notwithstanding the care taken to exclude it from the antenna and feeder system. For a discussion of this effect and its cure see Section 6, article 32.

WEATHER AND TIME TRANSMISSIONS. Radio navigational warnings and standard time signals are available to all ships equipped with communication receivers. The navigational warnings include the local weather forecast plus any urgent information with regard to tidal waves, offshore winds, ice, and storms.

10. RECENTLY INTRODUCED NAVIGATIONAL AIDS

RADAR. (See references 7–12.) Present radar systems provide one of the few known, yet most reliable, methods for surface-obstacle detection under conditions of restricted visibility, whether these obstacles are other ships, icebergs, buoys, islands, landmarks, etc. Radar provides, in effect, an electronic searchlight aboard ship, capable of "seeing" through darkness or fog, in any weather condition, for ranges up to approximately 25–30 miles (line of sight), depending on the power of the "searchlight." The use of radar, therefore, will be most applicable to collision prevention at sea, iceberg detection, harbor navigation, coastal navigation, and harbor control from shore stations.

Table 2 lists minimum specifications for marine navigational radar. These specifications suggest the use of PPI (plan position indication) presentation of the echo information (see Section 15, article 24).

Table 2. Performance Factors for Some Marine Navigational Radars

Performance Factors	U.S.C.G. Class A Spec.*	Manufacturer A	Manufacturer B	U.S.C.G. Class B Spec.†	Manufacturer C	Manufacturer D
Maximum range	30 miles	32 miles	40 miles	30 miles	30 miles	50 miles
Minimum range	100 yd	100 yd	100 yd	400 yd	200 yd	100 yd
Range resolution	100 yd	100 yd	100 yd	200 yd	100 yd	....
Bearing resolution	4° (on 10 cm), 3° (on 3 cm)	3°	3°	6°	3°	....
Presentation	7" PPI	7" PPI	12" PPI	7" PPI	7" PPI	7" PPI
Range scales (miles)	2–5, 4–15, and 15–30	2, 8, and 32	2, 10, and 40	2–5, 4–15, and 15–30	2, 6, and 30	1.5, 5, 15, and 50
Range accuracy	±2% or ±50 yd	....	±2%	±2% or ±100 yd	±2% or ±100 yd	±2%
Bearing indicator	True	Relative true (avail.)	True	True or relative	True	Relative true (avail.)
Antenna:						
Beam width	4° H, 15° V (on 10 cm); 2° H, 15° V (on 3 cm)	2° H, 15° V	2° H, 15° V	5° H, 15° V	5° H, 15° V	3.5° H, 12° V
Rotation	360°, 6–15 rpm	360°, 12 rpm	360°, 15 rpm	360°, 6–15 rpm	360°, 11 rpm	360°, 7 rpm
Frequency	3000–3246			3000–3246	3200	3070 ± 50
(Mc/s)	9320–9500	9320–9430	9320–9430	9320–9500		
R-f source	Magnetron	Magnetron	....	Magnetron	Magnetron	Magnetron
Peak power	15 kw	15 kw	35 kw	7 kw (on 10 cm), 15 kw (on 3 cm)	7 kw	15 kw
Pulse rate	800 cps	2000 cps	800 cps	800 pps	1500 pps	1000 pps
Pulse width	0.5 μs (maximum)	0.4 μs	0.25 μs	1.0 μs (maximum)	0.5 μs	0.4 μs
Receiver:						
Band width	Optimum	....	8 Mcs	Optimum	Optimum	3 Mcs
Gain	120 db	....	....	120 db	117 db	....

* Class A corresponds to U.S.C.G. Minimum Specifications, Brief No. 1, Nov. 14, 1945 (revised Aug. 1, 1946).
† Class B corresponds to U.S.C.G. Minimum Specifications, Brief No. 2, Nov. 14, 1945 (revised Aug. 1, 1946).

Principles. A short powerful burst of electromagnetic energy is emitted at a known spot and is narrowly beamed in a given known direction. Returning echoes from objects within an arbitrary range in that given direction are received at the known spot, detected, and visually displayed. For persistence and continuity of display, the emitted pulses are repeated periodically at a fixed rate with enough intervening time to allow the return of any echoes. If an echo from an object is received by the system after a time delay, T, from the initial burst, the distance of the object from the radar system is:

$$d = \frac{cT}{2} \tag{3}$$

where c is the velocity of light (or electromagnetic waves) in air and where the time T is measured from the beginning of the transmitted burst to the beginning of the received echo. Thus, the distance to the object is determined. If the antenna system is sufficiently directive, the pointed direction of the antenna at the time an echo is received is the direction of the echo, thus furnishing a bearing determination.

Unless the wavelength of the radiations used is small compared with the linear dimensions of the reflecting object, the phenomenon of diffraction takes place, making the echo amplitude inversely proportional to the square of the wavelength. If the wavelength is small compared with the reflecting object, the amplitude of the echo field does not sensibly depend upon the magnitude of the wavelength but rather upon the nature of the reflecting object. In the case of the free-space propagation between the radar system and the reflecting object, the following relationship holds:

$$P_r = \frac{A_0 G_0 P_t K}{(4\pi)^2 r^4} \quad \text{or} \quad r_{\text{max.}} = \sqrt[4]{\frac{A_0 G_0 P_t K}{P_{\text{min.}}(4\pi)^2}} \tag{4}$$

where P_r is the received power at the receiver input terminals (watts), A_0 is the effective absorption cross-section of the receiving antenna (square meters), G_0 is the overall power gain of the feeder and the radiating antenna, P_t is the transmitter peak power (watts), K is a complex reflection coefficient dependent on the nature of the target and is given as the effective echo area of the target in the direction of the radar (square meters), r is the distance in meters from the radar to the target. The second form of the equation is in terms of the maximum range, $r_{\text{max.}}$, and the minimum detectable power received, $P_{\text{min.}}$. These results for free space propagation must be modified for propagation over a spherical earth since, in the range of frequencies used, electromagnetic waves are propagated over approximately a "line of sight" path with small diffraction effects occurring at the lower radar frequencies. Thus, the antenna height above the sea, the target height above the sea, and the height of any intervening objects must be taken into account.

From these considerations, it is evident that, in order to utilize efficiently the principles of reflection with electromagnetic waves, a radar system must (1) generate a wave whose length is small compared with the objects from which the wave is to be reflected; (2) generate enough power at that frequency to be able to receive and detect the returning signal; (3) provide a means of measuring the time delay from the transmission to the reception of that signal.

Fundamental System Constants. Each radar system has associated with it certain constants whose choice depends upon the available components, the desired operational performance, and the intended use of the system. The normal variations of these constants are as indicated in the following:

(*a*) *Carrier Frequency.* The choice of carrier frequency depends on the permissible dimensions of the antenna system to be used and the directivity or beam sharpness desired, since the size of the antenna system reduces with increased frequency, and the directivity, as well as gain, improves with frequency for a given antenna size. Thus the lower frequency is limited by the practical antenna size; the upper frequency is limited by atmospheric reflection and absorption effects (pronounced near 3000 Mc or 10 cm wavelength for reflection and 30,000 Mc/s or 1 cm wavelength for absorption) and the availability of tubes capable of generating and amplifying enough radio-frequency energy to provide the necessary range. The lower frequency limit is about 100 Mc, though frequencies higher than 30,000 Mc have been successfully used experimentally.

(*b*) *Transmitter Pulse Width.* The minimum range at which an object can be detected by a radar system is determined largely by the width of the transmitted pulse (at the $^1/_2$ power point) since an echo returning while the transmitter is still operating will be masked by the transmitter pulse. This is even more true if the receiver is always blocked or desensitized for the duration of the transmitter pulse. The usual pulse widths range from 0.25 to 2.0 μs for navigational purposes.

(*c*) *Pulse Repetition Rate.* The pulse repetition rate must be slow enough to allow time for the maximum range echoes to return to the antenna before another pulse is transmitted, and it must be fast enough to provide enough traces while the antenna is rotating or pointing in a given direction to produce a lasting indication on an oscilloscope screen. Therefore the maximum range determines the highest pulse rate, and the rotational speed of the antenna determines the lowest pulse rate that can be used. In practice these rates vary from less than 60 to several thousand pulses per second.

(*d*) *Duty Cycle.* The duty cycle of a radar system is the ratio of the average power to the peak pulse power. It depends on the relation between pulse width and the pulse-repetition time. Thus a lower duty cycle permits higher peak power operation at the same average power. The maximum range of the system is dependent on the peak power for a given pulse width, while low average power means smaller tubes and components in the transmitter. The limit to peak power, however, for a given transmitter tube is the breakdown potentials between the various electrodes. In practice, radar duty cycles vary between 0.005 and 0.0001.

Fundamental System Components (General). Radar systems now in existence differ widely in detailed design and complexity, depending on their functional use and the accuracy and amount of information required. However, a single basic system can be visualized in which the functional requirements fit equally well almost all specific requirements.

The six primary components are shown in Fig. 5; they may be summarized as follows:

(*a*) *Timer or Synchronizer.* The timer determines the pulse-repetition rate of the radar and provides a zero reference point for time measurements and for operation of sequential functions in a definite time relationship. Such timing may be supplied by a separate unit such as a sine-wave oscillator, a multivibrator, or a blocking oscillator with the necessary pulse-shaping circuits. Another commonly used method of timing is to make the transmitter with its associated circuits establish its own repetition rate and provide the synchronizing pulses for the rest of the system. This action may be accomplished by a self-pulsing or blocking radio-frequency oscillator, or by a rotating spark gap. Self-timing eliminates a number of special timing circuits, but the pulse width and pulse-repetition rate are not usually as stable or rigidly controllable as is necessary for some applications.

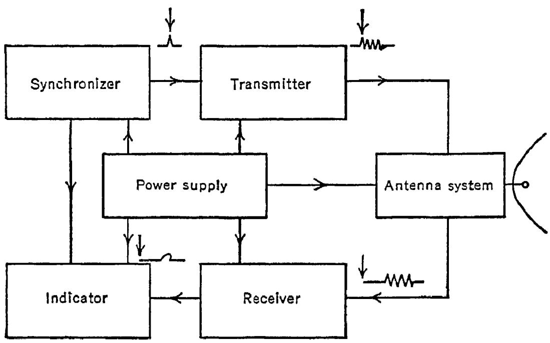


FIG. 5. Functional Block Diagram for Basic Radar System

(*b*) *Transmitters.* To generate pulses of high-frequency electromagnetic waves at high power levels a conventional pulse transmitter is used. Care must be taken that the tubes are suitable not only for the average power dissipation but also for the high powers and voltages during the pulse.

(*c*) *Antenna System.* The purposes of the antenna system are to beam and radiate the energy efficiently from the transmitter into space, to focus and pick up the returning echo and pass it on to the receiver. A transmit-receive switch (TR box) must be included to prevent the transmitter energy from harming the receiver.

(*d*) *Receiver.* A conventional wide-band receiver is used.

(*e*) *Indicator.* To display the detected pulses visually so that range, bearing, or elevation of any echo source, or combination of these, can be determined. See Section 15, article 24.

(*f*) *Power Supply.* A conventional well-regulated low-impedance power supply is used. Pulse-forming lines build up a charge during the inactive period to store energy to discharge in the pulse.

Performance Factors. (*a*) *Resolution.* The resolution of obstacles by a radar system will depend on the pulse width, the effective antenna beam width, the receiver band width, the frequency, and the stability of the entire system. For a typical commercial navigational radar in the 3-cm (10,000-Mc) band, having a receiver band width of 8 Mc/s, antenna beam width of 2° (to half power points), and a pulse width of $^1/_4$ μs, the resolution is 100 yards in range and 3° in azimuth. Above approximately 1000 Mc/s the frequency is not a major determining factor in itself, but only as it affects the other factors, since the wavelength is small enough to provide efficient reflection from the smallest objects ordinarily encountered.

(*b*) *Maximum and Minimum Range.* The maximum range will depend on the height of the antenna system above the sea, the power output of the transmitter, and the gain and efficiency of the antenna system. An average ship installation would have a maximum range of from 25–30 miles, which is, in general, great enough for the intended navigational use. The minimum range will depend directly on the pulse width and recovery time of the receiving system (including antenna switching mechanism); for the typical commercial system mentioned above, it is 100 yards. The radar indicator is usually provided with a range switch permitting any of several discrete ranges to be displayed on the screen. The accuracy of the range information depends on the initial accuracy and stability of the range markers; azimuthal accuracy usually depends on the accuracy of tracking between the antenna rotation and the sweep rotation on the cathode-ray tube face. Accuracies of 2 per cent in range and 2° in azimuth are not unusual.

(c) *Presentation.* The most usable form of the radar information for marine navigational purposes is the PPI type presentation on a screen of 5 to 9 in. in diameter, with provision for true or relative bearing stabilization, range markers, and range scale selector. The ratio between different range scales is often made the same.

(d) *Installation, Maintenance, and Operation.* The present radar systems are designed for operation by bridge personnel having little or no technical training. The indicator is mounted in the pilot house, and the antenna is properly located to provide 360° clearance to the horizon. To facilitate this arrangement on all types of vessels, the radio-frequency components, the antenna assembly, and the indicator are usually manufactured as separate units.

RADAR-BEACONS (RACON). (See references 10–12.) Radar piloting is beset by two important problems. First, targets for navigational purposes are sometimes small or poorly defined, resulting in small or weak echoes, especially near the maximum range of the radar system. Secondly, these weak echoes in the presence of sea return often leave the absolute identification of the navigational targets in doubt. These problems are greatly simplified, however, by providing a system of electronic beacons designed to serve as aids to navigation for ships equipped with the proper radar equipment. The term Racon, a contraction of *ra*dar-bea*con*, designates such a system. These beacons or racons are designed to emit or reflect a large amount of energy which will allow dependable target indication at a much greater distance than that obtainable when normal reflections alone are depended upon. There are three types of such beacons, and each is described below:

Radar Respondor-beacons (Transpondor). These racons consist of transpondors, or pulse-type receiver-transmitters, located at strategic ground sites such as coastal points and islands, which receive interrogating signals from a radar transmitter and automatically send back identifying reply pulses to the radar. The coded reply appears on the radar oscilloscope in such a way that both position and identification of the racon station are indicated. Reference to prepared code lists or charts then serves to associate that respondor-beacon with a fixed navigational point. Thus a navigator will be able to check his course, or even completely guide his craft, since range, bearing, and identification will be obtained from each respondor-beacon. A typical radar and fixed station microwave racon transpondor operation is shown in the block diagram in Fig. 6.

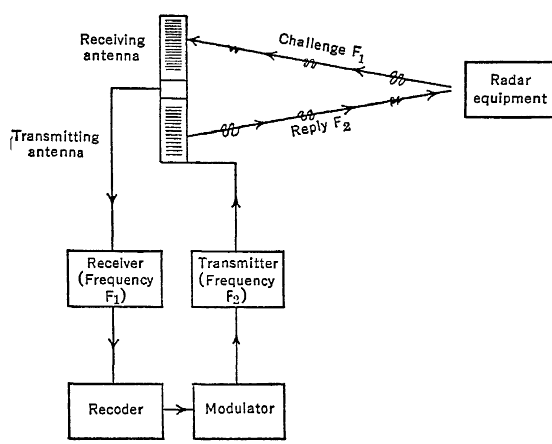


FIG. 6. Functional Block Diagram of Basic Racon System

To prevent racon signals from appearing on the radar screen all the time, and to avoid continual interrogation of the beacon by radar, the system is designed to operate under the following conditions: (1) the beacon responds only to pulses of 2-μs or greater duration within an appropriate frequency band; (2) the beacon replies at a fixed frequency which is common to the entire beacon system and which is just outside the radar system frequency band. Ten-centimeter band racons respond to 2-μs challenging pulses in the 3000–3246 Mc/s band and reply at 3256 Mc/s; 3-cm band beacons receive pulses in the 9320–9500 Mc/s band and reply at 9310 Mc/s. An additional band near 5 cm is being made available for radar and racon use. For racon system use, the radar equipment is designed to provide the necessary challenging pulse length and the correct receiver tuning upon activation of a special button or switch. The normal radar plot will then disappear and only radar-beacon responses will appear on the screen.

Three types of coding, to insure positive identification of a given respondor-beacon, have been devised: (1) Sequence coding is provided by emitting a number of half-second transmissions whenever the beacon is triggered. During the half second, pulses of a given duration appear as a pip of a certain width on the scope. However, the pulses making up the next transmission may all be of longer duration, thus giving a pip of greater width on the scope. The complete code is made up of the number and combination of widths of these pips. (2) Gap coding provides periodic interruption of a series of pulses. During each period of interruption, identification is given by a pulsed dot and dash transmission. Both sequence and gap coding are slow, requiring considerable time for complete identification. (3) Range coding gives an immediate identification as the complete code is produced on the oscilloscope at once. Each reply, caused by one interrogation pulse, consists of six pulses which appear on the PPI at different positions along a given bearing with the first pulse indicating the beacon position. Such coding is accomplished by changing the spacing of the pulses.

Separate broadside arrays serve as receiving and transmitting antennas for the transpondor. Both are non-directional in the horizontal plane in order to respond to interrogations from any point on the compass, but have a narrow low-angle vertical pattern to facilitate long-range operation. From the receiving antenna the received radar signals go to either a crystal-video or superheterodyne receiver which must satisfy the following requirements: (1) the radio- and intermediate-frequency stages must be broad-band in order to receive signals at any frequency within a given radar band (10-cm or 3-cm, but not both); (2) the video stages must not excessively widen 1-μs radar pulses; (3) the receiver must not be easily blocked; (4) the amplification must be adequate for the received radar signals. The output of the receiver goes to a recoder which consists of a discriminator, coder, and gate multivibrator. The discriminator accepts variable amplitude pulses of 2-μs or greater duration but rejects all pulses of shorter duration. The coder (in this case, a range coder) consists of six start-stop multivibrators whose outputs are combined in an amplifier to produce a series of pulses. The code may include a maximum of six pulses with 15- or 35-μs spacings between successive pulses. The gate consists of a 175-μs negative rectangular pulse that is fed back to the discriminator, where it prevents the coder from accepting other receiver video signals until after the completion of one range-coded transmission. The pulses from the coder are applied to a modulator where they are converted into very large amplitude rectangular pulses of $^1/_2$-μs duration. The modulator triggers a transmitting magnetron that oscillates at the proper reply frequency. To insure operation exactly on the radar transmission frequency, an automatic frequency control circuit is used to keep the magnetron in tune. Minimum peak powers of 10 kw are used, while some racons operate at peak powers as high as 40 kw.

Transpondor beacons of the type described, and radar equipments which use them, must be designed as "sister" equipments, as each is dependent upon the other. For this reason the Coast Guard's specifications for commercial radar equipment recommended that provisions be made for convenient future modifications of the radar for use with racons or microwave beacons. Considerable investigation and development work has been under way to permit adaptation of the wartime racon system to general marine navigational purposes.

RADAR MARKER BEACONS (RAMARK). A second type of radar beacon consists of a simple shore-based, continuously pulsed beacon operating on a single frequency. Rather than being interrogated by the radar pulses, this type continuously transmits, and the radar need have only a means for receiving the beacon frequency when it is desired to utilize the beacon. This eliminates the necessity for sending out 2-μs pulses and simplifies the ground equipment so that it acts in much the manner of the present low-frequency radiobeacons. The navigator sees, on his PPI, a line extending radially from the center to the periphery of the scope on the bearing of the beacon. Two beacons are needed to obtain a fix since no range information is provided. The means of identification is the location of the aid with respect to the naturally distinguishable land contours shown on the radar. Additional means of identification have been considered also, e.g., groups of pulses, known repetition rates, or modulation of the signal with an audio code, but these require additional modifications to the radar.

RADAR REFLECTOR BEACONS (CORNER REFLECTORS). A third type of beacon classed as a racon is the "corner reflector," which is merely a mechanical assembly of sheet-metal faces, designed to give the maximum reflection of energy received from any given direction over fairly wide angular limits. The reflector is inexpensive and easy to construct and maintain. These devices have no means of identification other than their prominent echoes on the radar screen and the fact that their positions can be plotted on charts. The great simplicity and economy of such a scheme suggest its use wherever feasible. In rough weather, however, it may be very difficult to distinguish its return through the ground clutter.

LORAN. (See references 11, 13, and 14.) The accelerated development of radar and other electronic equipments during World War II has reduced to relatively simple operational procedures the accurate measurement of radio transmission propagation time. This has resulted in the introduction and wide use of two important radio position-determining systems utilizing this propagation time principle, i.e., Loran and Gee. In these systems, a position fix is obtained from the intersection of two loci of position, each locus of position being determined by measuring the *propagation time difference* between two synchronized pulsed signal emissions arriving from two known but widely spaced radio transmitting sources. Since the velocity of propagation of radio waves over the earth's surface is essentially constant, propagation-time-difference measurement varies directly as the distance difference between a receiving station and two fixed transmitters. Therefore, the locus of points which are a given constant time difference from two fixed points is a hyperbola with the fixed points as foci. Thus the two transmitting stations or fixed points provide a family of hyperbolas about the stations, each hyperbola representing a constant value of propagation time difference. Position-fixing methods which use such a family of hyperbolas can be classed as hyperbolic systems, and, in addition to Loran and Gee, they include the Dingley system, Decca, and Popi. These last two systems utilize hyperbolas derived from constant propagation phase differences (identical in principle to optical interference phenomena) instead of constant propagation time difference, whereas the Dingley system utilizes hyperbolas derived from constant propagation frequency differences.

STANDARD LORAN. This system operates at approximately 2 Mc/s and utilizes ground wave and/or single hop *E* layer paths of transmission. Special charts are provided having Loran hyperbolas superimposed on geographical maps. Figure 7 shows a partial example of Loran hyperbolic charts. In practice, a master (or A) station transmits 40-μs pulses at a given repetition rate. These pulses serve to synchronize a slave (or B) station several hundred miles away so that the B station also transmits 40-μs pulses at the same repetition rate, but delayed in time by the travel-time between the two stations plus the B station's system delay. The reception and presentation of these transmitted pulses aboard ship by means of a special Loran receiver and indicator permit measurement of their time difference, and from that information the appropriate Loran chart will give the correct corresponding hyperbolic line of position. A separate operation on a second pair of Loran transmitting stations provides a second position curve, and the intersection of these hyperbolic curves on the special charts establishes a "fix." A block diagram of the Loran system is shown in Fig. 8.

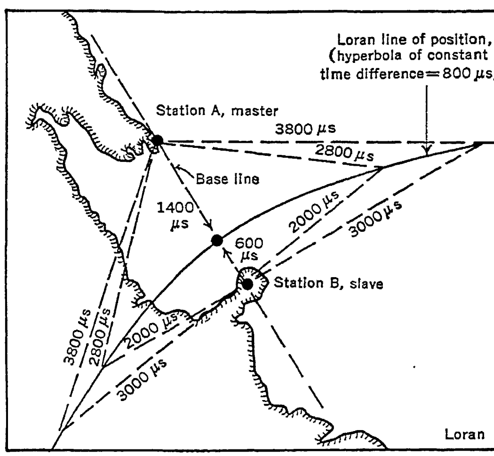


FIG. 7. Loran Hyperbolic Line of Position

Transmitting Equipment. The Loran transmitting stations generally operate as a group of four stations. Each station would normally consist of a timer and a transmitter. The timer establishes a very precise pulse rate by utilizing 50 kc/s or 100 kc/s crystal-controlled oscillations of very high accuracy, which are divided down to provide basic pulse rates of 25 pps and $33^1/_3$ pps. These timing pulses trigger an exciter which generates the 40-μs pulses for modulating the transmitter. The transmitter itself is a self-excited, tuned-grid, tuned-plate, push-pull unit, pulse modulated in the cathode circuit and oscillating from 1750 to 2000 kc/s. To provide precise synchronization, to conserve equipment, and to reduce maintenance, a single transmitter is usually employed as a master station which synchronizes two different slave stations, on the same frequency, but at different repetition rates. This is accomplished by providing, at the transmitter, two separate timers (each of which is independently referenced, for convenience, to the precise time signals from station WWV), two exciters, and a mixer stage preceding the modulator. The duty cycle of this double-pulsed transmitter system is approximately 80 μs in every 40,000 μs (25 pps), or 0.2 per cent. The transmitter, operating at about 100 kw peak power, feeds a 120-ft guyed radiator in the standard installation. In case of damage by windstorm, etc., an emergency T antenna is provided. The slave-station emitting equipment is identical to that of the master station, and suitable means are provided at the slave station for properly receiving the master station signals and for effecting the necessary precise synchronization of the slave-station pulses.

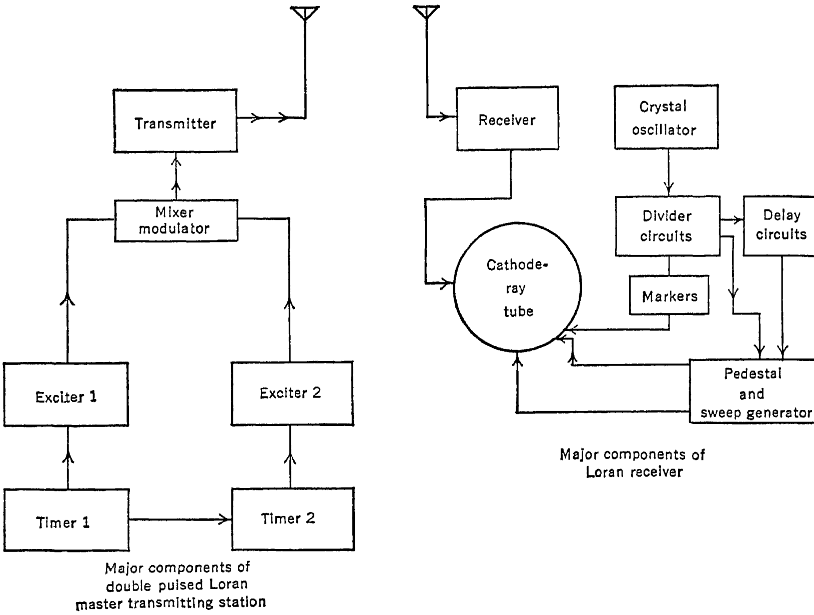


FIG. 8. Functional Block Diagram of Basic Loran System

RECEIVING EQUIPMENT. The craft's receiver-indicator equipment is usually furnished with a 50- to 60-ft vertical antenna leading to a medium-frequency receiver having a band width of 50 kc/s to pass the 40-μs pulses, and a 50-kc video output stage. To insure maximum range reception, the receiver must have enough sensitivity and gain to provide full screen deflection with 10 microvolts into the receiver. The indicator provides the necessary timing circuits to measure pulse separations with the required 1-μs precision. A 100 kc/s precision oscillator with associated dividers supplies a sequence of precise index timing markers at appropriate intervals, with a basic accuracy of ± 1 μs. It also contains the delay and deflection circuits for displaying the received pulses or the timing markers on the oscilloscope screen. The actual method of presentation is to provide separate paralleled sweep traces on the scope screen, one for the train of A station pulses and one for the train of B station pulses, with the index markers appearing simultaneously on the screen with the station pulses. Since more than one Loran transmitter may be operating on the same carrier frequency, pulses from all these stations will appear on the indicator screen. Therefore, the indicator includes timing and sweep circuits which provide precise repetition rates corresponding to the desired transmitter's repetition rates. This procedure synchronizes the indicator to the pulses of only one transmitting pair of Loran stations while the pulses of all others continuously drift across the screen and do not interfere with the desired stationary pair of pulses.

The exact amount of time delay between the slave and master pulses can be measured from manually incorporated time delay circuits required to make both these pulses coincide on the screen. The accuracy of measurement depends on the stability of the circuits which provide timing markers on the face of the oscilloscope. To facilitate this, an expanded or fast sweep is provided, enlarging the picture of the pulses and making the pulse-matching procedure accurate to approximately 1-μs, since the 10-μs marker spaces can be estimated to tenths. A typical presentation is shown in Figs. 9(*a*) and 9(*b*). A receiver refinement introduced by the Sperry Gyroscope Company permits the local time delay (equivalent to the time difference) to be read directly in microseconds on a Veeder-type counter. The currently available shipboard receiver-indicator equipment weighs between 125 and 235 lb, although the corresponding airborne equipment is a comparatively small unit weighing approximately 25 lb. Both types consume 200 to 300 watts from a 115-volt power source.

Performance. When used over water Loran can provide navigational fixes up to 750 nautical miles in the daytime and up to 1400 nautical miles at night. The daytime range is the limit of ground-wave propagation at the frequencies and power used. At night, pronounced reflection effects from the E layer of the ionosphere provides the additional range. Fortunately, this ionospheric reflection introduces only small errors which are compensated by known correction factors and allowed for in the Loran charts. When the ship is in a position to receive both ground- and sky-wave transmission, the indicator will show two or more sets of pulses where the first set is the ground-wave pair and the others are various sky-wave reflections as shown in Fig. 9(a). Such reception does not interfere with the fix determinations. The average operating accuracy of present systems is better than 0.5 per cent of the distance measured, e.g., less than $2^1/_3$ miles in 500 miles, but deteriorates somewhat along the base line of the transmitters. The accuracy depends on the location of the transmitting stations, the stability of the timing circuits, and the experience of the operator in matching pulses and estimating time between 10-μs markers.

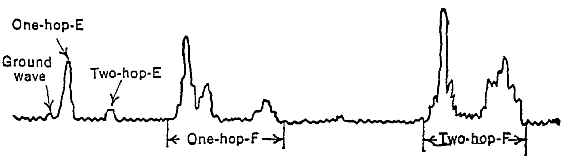


Fig. 9(a). Typical Ground Wave and Sky Wave Pulse Sequence from One Station. Signals to the right of the One-hop-E are generally disregarded.

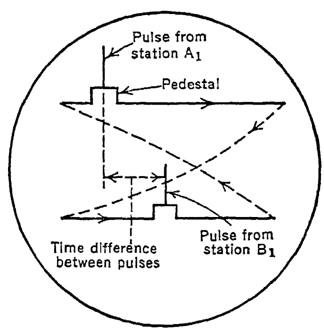


Fig. 9(b). Basic Loran Indicator Pattern. A calibrated time difference (or delay) mechanism and an expanded sweep are used to obtain 1 microsecond accuracy.

Since the system depends upon the measurement of time differences rather than upon direction of propagation, it is not subject to the usual errors encountered in direction-finder systems. Factors affecting the propagation from a master station will, in general, also affect the propagation from its slave station. This is especially true as the range *increases*, since the two paths then become approximately parallel. At the close of the war, the military Loran system comprised 40 transmitting stations located at strategic geographical points, with an additional 10 under construction; these systems supplied information to receiver-indicators installed on more than 3000 surface vessels. Among the improvements which are desired and may be expected in the near future are the following: (1) higher-power transmitting stations (about 1000 kw), (2) crystal-controlled oscillators to provide more stable carrier frequencies, (3) different methods of station synchronization and pulse-shaping circuits, and (4) a totally automatic receiver-indicator system requiring a minimum of technical operating skill and experience. These improvements are desired in order to overcome the present disadvantages of: (1) limited range, especially in the daytime, (2) the undesired synchronization upsets caused by ionospheric disturbances, (3) the need of a specially trained operator for interpretation and maintenance. In addition, experimental studies are being conducted (see below) on LF Loran which is an adaptation of standard Loran to the very low-frequency range.

SS LORAN. This system (*s*ky-wave *s*ynchronized Loran) is operable during the night only and is identical to standard Loran except that master and slave stations are separated by 1000 to 1200 nautical miles instead of the usual 200 to 300 nautical miles. Such a long base line requires synchronization of master and slave stations by sky waves, since, at the operating frequency of 2 Mc/s, the range of the ground-wave propagation is less than the base line itself, so that SS Loran can be used only at night, when sky waves are strong.

LOW-FREQUENCY LORAN. To provide greater daytime range, a modified Loran system has been designed which operates at about 180 kc/s and offers the following propagation advantages: (1) at these low frequencies the propagation range of ground waves is increased, especially over land; and (2) sky-wave reflections become much more stable, to such a degree that they can more readily be interpreted by a navigator. This stability results from the fact that radio-frequency energy at these frequencies does not penetrate the E layer of the ionosphere and thus the long trains of multiple night time sky waves, present at 2 Mc, are absent. Hence, the 180 kc/s sky-wave receptions are much more useful to the operation of the Loran system and it appears possible to obtain ranges of 1000 or more miles for satisfactory, stable operation for day or night over land or sea. However, resolution (separation) between the ground-wave and various sky-wave receptions deteriorates and affects the accuracy of the system. At very low frequencies (below 180 kc/s), ground waves are even less attenuated, and very reliable propagation over thousands of miles can be expected. However, because very wide-band antenna systems are not available at the very low frequencies in view of their prohibitive size and because of current limitations of wide-band and pulsed techniques, the pulse principles inherently required in Loran have not yet been adapted to the very low frequencies employed. In view of these factors it seems that 100–150 kc/s is about the lowest usable frequency range in which pulsed systems such as Loran can be employed, with the pulse length increasing to about 300 μs.

GEE SYSTEM. The British Gee system of navigation operates in the 20–85 Mc/s frequency range, on the pulse-propagation-time-difference principle. In nearly all respects it is the higher-frequency counterpart of the Loran system and was independently developed by the British. The essential differences of the Gee system are: (1) higher operating frequency and consequent reduction of range; (2) three-slave-station operation instead of two; and (3) simultaneous presentation of two propagation time differences in order that a position fix can be determined in one initial receiver operation.

CONSOL (BRITISH) AND SONNE (GERMAN). (See references 15 and 16.) Consol or Sonne is proposed as a medium-range navigational aid of the rotating beacon class. Radio navigational aids of this class operate in much the same manner as a directional searchlight which revolves at a known and constant rate with the addition of a non-directional visual light signal marking the instant that the revolving light passes through true north. The observer, at the point where the line of bearing to the beacon is to be determined, needs only to employ a stop watch to measure the time interval between the reception of the non-directional light signal and the searchlight signal, after which the desired azimuth angle is computed from the product of the known angular rotational speed and the time interval. In the Consol system dots and dashes are put into the rotating antenna pattern so that a stop watch is not needed. The observer simply counts the number of dots and dashes between the "north" pulse and the equisignal, and obtains his azimuth from the station thereby. A variation of this rotating-beacon principle consists of directly imparting the desired azimuthal information to the light beam proper via correlated modulation methods, e.g., by coding, intensity modulation, or frequency modulation. Completely analogous rotating *radio* beacon systems have also been used. For low and medium frequencies, large rotating antennas, or fixed antennas (e.g., Adcock type) with rotating goniometer arrangements transmitting a cosine-law or cardioid pattern, have been used. In other systems a cardioid or cosine-law modulation pattern has been transmitted, utilizing a receiver with a pattern synchronized cathode-ray sweep for an automatic indicator. Generally, radio frequencies in the VHF and UHF regions have been favored because rotating antennas of reasonable dimensions can be more conveniently built to give the desired sharpness of beam width to the predominantly single lobe radiation pattern usually employed. The necessarily high carrier frequency of these systems unfortunately limits them to line-of-sight ranges. Consol or Sonne, on the other hand, is a rotating-beacon system that operates in the 200–500 kc/s band and provides an accurate medium-range facility. Use is made of an ingeniously employed rotating multilobe radiation pattern such that sharp bearing resolutions are realized at these ranges.

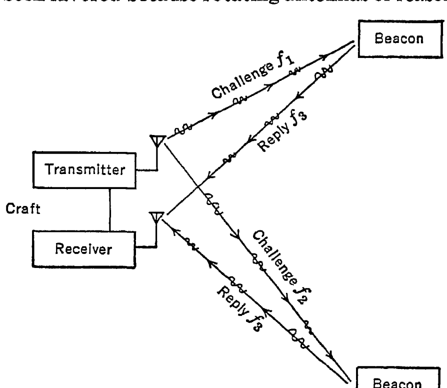


FIG. 10. Functional Block Diagram of Basic Shoran System

SHORAN. (See reference 17.) The Shoran, or *sho*rt *r*ange *n*avigation, system operates on the principle of echo ranging on two spaced beacons of the radar-respondor type. It combines the accuracy of propagation-time measurement with simplicity of operation and reliability of equipments.

Briefly, Shoran position "fixing" requires the measurement of the round-trip transmission times, in terms of range or distance, of two sets of short radio-frequency pulses transmitted by the craft, each set interrogating a known shore radar respondor-beacon station which retransmits the pulses back to the craft. Since each echo range defines an arc of fixed radius from each known radar respondor-beacon position, the intersection of these arcs determines the craft's location. The resulting double ambiguity, as in Loran and Gee, can usually be easily resolved. The two sets of pulses are transmitted from the craft on separate frequencies as shown in Fig. 10. Upon interrogation, both ground beacon stations retransmit on a common frequency that is different from either of the signal frequencies radiated by the craft. The returned pulses are then received and displayed by the craft's equipment. The crystal-controlled repetition rate is the precise yardstick in terms of which the craft to respondor-beacon distances are measured, and all interrogating transmitters use this same repetition rate.

DIRECTION-FINDER SYSTEMS (HF-VHF-UHF). (See references 18, 19, and 20.) Until recently direction-finder systems at these frequencies were quite unreliable. Primarily, this poor operation has been caused by resonant reradiation from structures near the DF collector. Since a grounded metallic structure will resonate approximately as follows:

$$f_{(\mathrm{Mc/s})} = \frac{246}{(h)_{\mathrm{ft}}} \tag{5}$$

structures between about 8 ft and 165 ft high will have a resonance somewhere in the 1.5 to 30 Mc/s frequency band. (See Environmental Effects, p. 22-39.) On the basis of these analyses shipboard HF/DF can now be installed with average errors of 5–10°.

In view of their poor accuracy compared to the 1° to 2° shipboard MF/DF accuracy, it is not likely that shipboard HF/DF will be widely employed for other than military intercept and location applications.

Shore stations (see reference 20) of 100 Mc/s direction finders employing a vertical spaced Adcock collector, a spinning goniometer, and a cathode-ray-type automatic indicator have given a 2–3° accuracy with line-of-sight range. At 8 Mc/s planes have been tracked from the United States to Northern Africa. For the ultra high frequencies 90 to 5000 Mc/s systems use a spinning reflector-type collector (dish) system in lieu of the fixed Adcock and goniometer arrangement. In general these are not too efficient in the 90 to, say, 200 Mc/s range but rapidly improve for the higher frequencies.

SONAR. (See references 21 and 22.) Sonar, or *so*und *na*vigation and *r*anging, which is the field of underwater sound navigation, has been applied in a manner similar, in principle and in electronic circuits, to radar in the field of radio navigation; the principal difference is the medium of propagation. The two main types of underwater sound aids in use are: (1) the vertical projectors called echo sounding or depth finding aids; and (2) the horizontal projectors known as echo ranging units.

Echo sounding equipment transmits sound pulses vertically downward from a projector placed in the bottom of the ship, or lowered from the side of the ship, and receives the reflected pulses from the bottom of the ocean or waterway, giving a direct and constant running record on graph paper of the depth of water beneath the boat's keel. Present commercial equipment has a range of 1000 fathoms or less and is used to navigate channels, rivers, and other waterways and to obtain data for charting the depths of these various waterways. Information necessary for the plotting of charts having contour lines of constant depth is gathered by the Coast and Geodetic Survey and is published by the Navy Hydrographic Office as an aid to marine navigation. This type of equipment has also been successfully used by fishing boats to locate schools of fish, as they also reflect sound pulses.

Echo ranging equipment transmits sound in a pattern which is conically beamed. It searches in the horizontal plane from a projector which can be rotated through 360°. Originally, the term Sonar implied the use of ranging equipment only, but it now designates almost any type of underwater sound equipment regardless of its use. This equipment was originally developed for the U. S. Navy for detecting submarines and surface ships, but it can obviously be used in the same manner for navigational and anticollision purposes. It presents the distance from the ship to an object, directly in yards, on a cathode-ray range indicator, and the bearing of the object with respect to the ship's position on a bearing indicator compass.

11. CONTEMPLATED AND PROPOSED NAVIGATIONAL AIDS

LANAC. The Lanac or *la*minar *na*vigation and *a*nti*c*ollision system is a method of electronic navigation, developed since the close of the war, which proposes a unified radar and identification navigational system for both marine and air use (see article 6), utilizing a minimum of equipments. The Lanac system provides target identification, coded radar beacon, and radar search operations from the same basic equipment. Essentially, the proposed system comprises a challenger and a replier, operating at approximately 1000 Mc/s, as shown in the block diagram Fig. 11. The challenger can also be operated as a low-power search radar, for auxiliary navigation and anticollision protection. The repliers are coded for identification and (in air use) for altitude discrimination. Thus the system provides the essential advantages of radar navigation, but the technical and operational emphasis is placed on the beacon operation, which provides positive point identification of all strategic locations and other ships, assuming that the ships are properly equipped. Thus, for maximum efficiency of operation, most other ships and prominent geographic locations would have to be marked by responder-beacons (repliers). The display of beacon replies on a PPI scope as a two-dimensional polar plot of the area surrounding the challenger is easier and more reliably interpreted than a corresponding radar display.

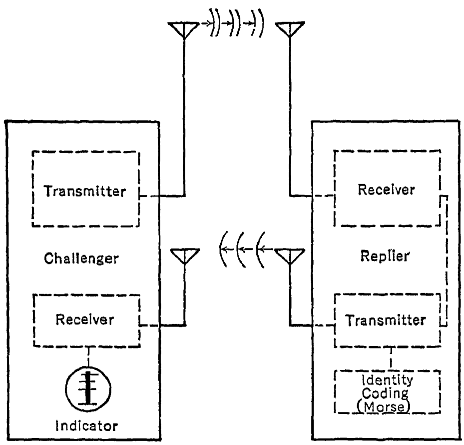


FIG. 11. Functional Block Diagram of Basic Lanac System

A chart of the marine services rendered by Lanac is shown in Fig. 12 with the addition of the auxiliary radar functions mentioned. Although not yet in general use, Lanac offers the advantages of a simple and very comprehensive navigational aid which combines the reliability of beacon operation with the versatility of radar.

DECCA. (See reference 16.) The British Decca system operates in the low- and medium-frequency ranges (i.e., 10–200 kc/s) on the propagation phase difference principle (e.g., similar to POPI). Operation is accomplished by transmitting CW signals from a master and a remote slave station accurately synchronized so that there is set up in space a stable interference pattern of radiations having distinct loci of constant phase difference.

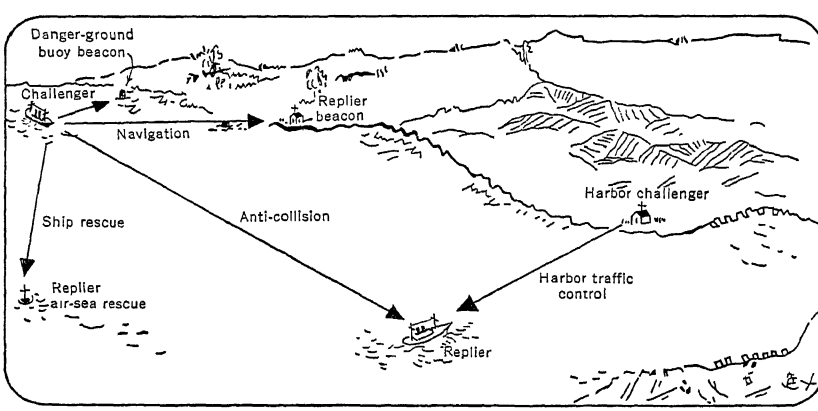


FIG. 12. Summary of Services Rendered by Lanac (Excluding Radar)

These Decca loci are hyperbolas of constant propagation phase difference, as compared to the Loran loci which are hyperbolas of constant propagation time difference. To facilitate the reception of the two distinct transmitted signals, and the practical measurement of their phase difference, the Decca master and slave CW emissions have a highly stabilized ratio of carrier frequencies; e.g., the master station may operate on $340/4 = 85$ kc/s, the slave station on $340/3 = 113.33$ kc/s. The received frequency ratio is converted to unity by suitable multipliers in two channels of the receiving equipment. The actual phase difference between the two signals is then measured by passing the two converted signals through phase discriminating and indicating circuits. An integrating phase-meter (called a decometer) numerically identifies each hyperbola of 0° phase difference as it is picked up by the receiver. These 0° phase difference hyperbolas are called "equiphase" lines and are used as a reference. All other hyperbolas within the "lanes" between the equiphase lines are then identified on the decometer as values of phase difference other than 0°. Since two such position lines are required for a "fix," another such phase comparison determination is necessary on a second pair of stations. The total equipment, therefore, comprises three shore transmitters (one master operating two slaves) together with their associated control circuits, and craft equipment consisting of three phase-stable amplifiers, four frequency-multipliers, and two integrating phase-meters.

In order to avoid ambiguities of fix in the Decca system, the decometers must be preset to a given reading at a known point which is within reception range of the Decca transmitters. Then, as long as the transmitters and the receiving equipments are operating, the decometer will read correctly as the receiver travels from one lane to another. If either the receiver or any of the transmitters fail to operate properly, or the receiver travels out of reception range, the absolute setting is lost and must be reset at a known position. To reduce these possibilities of ambiguity, means for positive lane identification are now being developed. A Decca facility (utilizing four stations centered near London, England) is in continuous 24-hour operation. Other Decca facilities are being planned including some for the European continent. The Decca system is easier to operate than the Loran system and appears capable of giving fix information with greater accuracy, but at decreased range. However, errors due to serious phase shifts may be introduced by ionosphere reflections of sky wave. It is to be noted that three carrier frequency channels are required for a position fix with the Decca system.

POPI. (See reference 16.) The British POPI or *p*ost *o*ffice *p*osition *i*ndicator system, operates in the medium-frequency range on the propagation-phase-difference principle (similar to Decca) with such a short base line that the hyperbolic lines of position are essentially straight lines or bearing lines of position to each POPI station. Two or more bearing lines of position from two or more POPI stations give the desired position fix. Each POPI station employs three outer antennas symmetrically disposed about a central antenna. If the antenna spacings are equal to or less than $\lambda/2$, there are no ambiguities, and at a distance exceeding five times the antenna spacing the hyperbolic lines may be considered as straight bearing lines of position. In operation, the central antenna transmits continuously at a radio frequency of f_3, the outer antennas transmit at a slightly lower radio frequency f_1, and they are sequentially keyed at some fraction, $1/n$, of the frequency difference f_2. For a typical POPI station, f_1 may be equal to 750.000 kc/s, $f_2 = 80$ cps, $f_3 = 750.080$ kc/s (obtained by mixing f_1 with f_2), and $n = 4$. With $n = 4$, the keying cycle would occur as follows:

First outer antenna on first $1/20$ sec
Second outer antenna on second $1/20$ sec
Third outer antenna on third $1/20$ sec
Space and reference on fourth $1/20$ sec

At the receiver, the audio beat frequency $f_2 = f_3 - f_1$ (e.g., 80 cps) is applied to a rotating switch with four contact sectors. In order to control this rotating switch so that the sequential transmissions will be correctly received, one of the 80-cps outputs from the rotating switch is used to synchronize an 80-cps oscillator which, in turn, drives a synchronous motor. This motor, through reduction gearing proportional to the factor $1/n$, permits $1/20$-sec intervals for receiving the sequential 80-cycle tones corresponding to the frequency difference of the transmissions from the outer and central antennas of the POPI station. Since each of the outer antennas has a separate distance or propagation time to the receiver, there is a correspondingly separate and distinct phase for each 80-cycle tone sector of the synchronous commutator's output.

It is planned to measure these phase differences as if they were coexistent even though they actually occur in alternate time sequence. One means proposed for accomplishing this measurement consists of a separate 80-cycle filter network for each synchronous commutator segment which is capable of ringing or sustaining 80-cycle oscillations after being energized for about $1/20$ sec without exhibiting adverse relative phase shifts between networks. No report is available at present on results obtained with this critical ringing network. The present British developments are in an experimental state, although full-scale trials are contemplated at an early date. It is to be noted that POPI may suffer the same errors, due to ionosphere phase shifts, that are expected in the Decca system.

TELERAN. (See reference 16.) Teleran, or *tele*vision *r*adar *a*nd *n*avigation, is a system that was originally proposed as an aid to air navigation and is discussed in detail in article 6.

It is believed, however, that it has considerable parallel usefulness as a harbor traffic control type of marine navigational aid, particularly during adverse weather conditions. Basically Teleran is a television system whereby the central control (shore) station transmits, by television, any essential information to nearby craft. In this manner pertinent PPI (plan position indication) information, weather warnings, and special instructions can be instantly presented via the television link to the navigator, using only a standard television receiver as the shipboard equipment.

FACSIMILE. Because of the increased speed with which navigational information will be available to the navigator in the future, it is expected that some storing and recording means will eventually be required in order properly to handle navigational information that cannot be too well interpreted via oscilloscopes, etc., under busy harbor and related conditions. Moreover, the need and desire for rapidly printed records of weather, warnings, and congested traffic conditions will serve as a powerful stimulus in the eventual adoption of such an aid. In view of the relative ease with which facsimile signals can be multiplexed with other transmissions to the craft being navigated, its introduction is not expected to be long delayed. In some respects, facsimile transmission becomes a desirable compromise between the systems of telephoto, television, and teletype transmissions. Details of the principles of facsimile are discussed in Section 19.

SOFAR. Sofar, or *so*und *f*ixing *a*nd *r*anging, is a proposed system for air-sea rescue service and is a possible long-range navigational aid in the field of underwater sound. It has been discovered that sound originating at certain depths (often 3000 to 4000 ft) will be refracted downward by the layers of water above the critical depth, owing mainly to the temperature gradient in the water, and upward by the layers of water below, owing largely to the hydrostatic pressure gradient, thus confining the sound energy to a horizontal channel. It has been found possible to use this channel for the transmission of a distress signal (e.g., by detonation of a suitable explosive charge) over a range exceeding 2500 miles. Reception of this signal by several coordinated and widely spaced hydrophone receivers allows an accurate position fix of the distressed craft. Special light-weight and properly armed charges have been designed for airplane and life-raft applications that are particularly efficient for air-sea rescue purposes. An initial network of four stations, at Hilo and Kaneoke in the Hawaiian Islands, and at Monterey and Point Arena on the California coast, principally for air-sea rescue purposes, cover the long air route from San Francisco to Hawaii.

REDAR. Redar, or *re*d (and infrared) *d*etection *a*nd *r*anging, is a proposed aid to marine navigation, similar in principle to radar. In redar, it is proposed to use a searchlight and telescope arrangement in the near-infrared band aboard the ship. The general techniques and procedures employed in radar are expected to be modified as necessary and applied to redar. For navigational applications, where rapid redar scanning is not necessary, a simplified infrared optical range finder operating on well-known optical range finding principles is proposed. At the present state of infrared investigations, ranges of 3 to 5 miles are predicted during average to clear weather and low humidity conditions, with accuracies comparable to radar accuracies. Because of the nearly equivalent attenuations of infrared radiations as compared to visible light radiations under fog conditions, infrared systems are not expected to become a general navigation system. If the need for very short-range target and obstacle detection becomes pronounced, it is possible that redar may adequately serve this need in a manner superior to that of most standard radar systems.

MISCELLANEOUS SYSTEMS. Dingley system (see reference 23) is a frequency-modulation type of radio navigation aid that basically operates on the propagation-time-difference principle, wherein the actual time difference is obtained in the form of a frequency difference (or audio beat). This is accomplished by the use of frequency-modulation emissions from two spaced antennas, whereby there is set up in space a stable pattern of frequency differences having distinct loci of constant frequency difference. Thus the Dingley loci are hyperbolas of constant frequency difference as compared to the Loran loci or hyperbolas of constant propagation time difference. In operation, a central antenna and at least one other outer antenna are energized with frequency-modulation signals. The signal received along a radial line from either antenna has a space distribution of frequencies when that antenna is emitting frequency f_0 at time t_0, f_1 at time t_1, f_2 at time t_2, etc. At time t_1 the signal of frequency f_0 has traveled to d_1, and signal of frequency f_1 is being emitted at antenna A or at d_0. The partial table shows the overall conditions responsible for the space distribution of frequencies. Since the table is also applied to give similar space distribution of frequencies for similar emissions from the second antenna, it can be seen that those points in space which maintain a constant distance difference to the two antennas will define a hyperbola along which will be found a constant frequency difference. Hence, a family of frequency-difference hyperbolas exists for the various differential distances to the two antennas, and their general similarity to Loran hyperbolas of constant propagation time difference is immediately apparent. If desired, the reference antenna, A, can be designated as master and any other antennas, B, C, etc., can be designated as slaves.

Instantaneous Time	Instantaneous Frequency at d_0	Instantaneous Frequency at d_1	Instantaneous Frequency at d_2	Instantaneous Frequency at d_3	Instantaneous Frequency at d_n
t_0	f_0	....	....	....	..
t_1	f_1	f_0	....	....	..
t_2	f_2	f_1	f_0	....	..
t_3	f_3	f_2	f_1	f_0	..
.	.	.	.	.	.
.	.	.	.	.	.
t_n	f_n	f_{n-1}	f_{n-2}	f_{n-3}	..

In view of the above, the Dingley system may be considered a simplified type of Loran system operating on an F-M basis. Because of the wide band factors generally associated with frequency modulators, it may be expected that the higher carrier frequencies may be required with their correspondingly reduced range.

Sonic direction finders have long been employed by the military particularly for anti-aircraft purposes. Similar sonic direction finders have been employed for underwater applications. In such applications, a pair of spaced sonic pickup devices (e.g., crystals acting as hydrophones or sound microphones) have their outputs differentially connected to yield a cosine law of coupling (i.e., similar to a pair of Adcock dipoles) and have their resultant output fed to a suitable receiver. When used with a system of underwater sound beacons the combined arrangement is similar to the radiobeacon system described under Established Navigational Aids, p. 22-36. Beacon stations are under investigation for automatically taking sonic bearings on a ship and transmitting by radio the range and direction of the ship with respect to the beacon's position.

Infrared systems have been proposed which are essentially similar to redar with an optional infrared beacon system on shore to provide a convenient navigational aid. The infrared systems, as pointed out previously, do not permit navigating during fog conditions owing to their inability to penetrate the fog without an attenuation almost as great as that experienced by visible light. Their chief value would appear to be a black light navigating system operating under blackout conditions and during the absence of fog. Such systems generally operate in the *near infrared* region (0.3 to 1.5 microns).

Far infrared (heat-ray) systems have been developed to distinguish targets from their backgrounds by operating on heat or temperature differences. These systems generally function in the 8- to 15-micron band and usually operate in a manner wherein the heat rays are detected by bolometers (see reference 24). One type of bolometer successfully developed by Johns Hopkins University operates as a superconductor at temperatures near absolute zero. This superconducting device may include scanning arrangements whereby target resolution, or definition, roughly comparable to radar definition, can be attained. Because of atmospheric absorption, ranges beyond a few miles are seldom obtained.

Microwave thermal detection has been reported by G. C. Southworth (reference 25) in an interesting series of measurements of microwave radiations from the sun. This phenomenon is to be expected since radio waves may be considered as infrared radiation of very long wavelength, and a hot body would be expected to radiate microwave energy thermally. Using such techniques, absorption bands of water vapor have been measured. It is to be further expected that targets casting a shadow by blanking out radiations over certain areas might be detected by such microwave thermal and similar techniques. Any radiation source or any radiation field arranged to cause target shadows is potentially a means for detecting these targets, whether the frequency of these radiations is sonic or extends to the cosmic-ray region, and whether or not the emitting source is under control of the local observer.

12. DETERMINATION OF OPTIMUM TRANSMISSION PARAMETERS FOR SOME LONG-RANGE RADIO NAVIGATION SYSTEMS

BASIC CONSIDERATIONS. A long-range transoceanic radio navigation system for global application must meet three essential requirements, namely, provide: (1) adequate signal reception with a sufficient number of coastal or island stations for world-wide coverage; (2) reliable signal reception at useful levels irrespective of weather, time of day, season, year, direction, and distance of reception; and (3) an economically feasible arrangement avoiding prohibitive installation and operating costs. P. R. Adams and R. I. Colin (reference 26) have studied these essential requirements and have arrived at the conclusion that, if the craft should at all times be within range of two ground stations (in order to obtain a cross fix), then these stations should have a minimum range of 1500 miles in order to insure world-wide double coverage. In addition, they have analyzed the relative suitability of the various transmission parameters for a reliable 1500-mile range of a long-range, essentially CW, radio navigation system. These analyses cover the optimum choice of frequency, band width, power, modulation, and radiation (antenna) efficiency, and comprehensively consider these transmission parameters with due respect to the following: (1) Quasi-minimum field strength (defined as the monthly average of signal strength measured at the worst hour of the day during the worst months of the worst year) that may be expected at different places, frequencies, direction of transmissions, etc. (2) Quasi-maximum significant static intensity (defined as the value of static which is rarely exceeded at any time and place where the useful signal is likely to have its quasi-minimum field strength) that may be expected at different places on different frequencies, etc. (3) Signal-to-noise ratio, based on the ratio of quasi-minimum field strength to quasi-maximum significant static intensity, in order to determine the ratio's minimum probable value, by giving proper regard to each factor in the ratio in order to avoid their extreme values when the time and place of their occurrence may never coincide (the time relationship for example, depends upon factors such as latitude, time of day, and direction of transmission). (4) Cost of reliability in the choice of a suitable transmitting frequency (aside from the normal merit factors determining a choice of transmitting frequency based on average reception conditions), the degree and frequency of occurrence of fadeouts, adverse polarization changes, static conditions, and susceptibility to magnetic storms. (5) Corona factors, since the size of antenna is partly determined by the required antenna capacitance that will prevent the antenna charging current from developing a peak voltage in excess of the critical corona voltage; the latter, in turn, is dependent upon the weather as well as the dimensional and electrical factors.

As a result of the above types of detailed investigation, employing compiled data and consideration of the pertinent factors including signal and static strength and fluctuations, and antenna efficiency, the conclusion is reached that, in an essentially CW type of transmitting system for navigational purposes up to a maximum distance of 1500 miles, adequate reception can be assured with least power input at a transmission frequency of 70 kc/s. These investigations show that the reduction in time lost through fading and fadeouts offsets the disadvantage of low-frequency transmission and its associated lower radiation efficiency and higher static intensity. Estimates of power requirements for several locations and antenna dimensions for handling such power without corona are also discussed in the above reference.

MODULATION AND BAND WIDTH. The above study makes little mention of the various methods of modulation by which intelligence may be transmitted at the chosen carrier frequency. Three paramount points must be considered: (1) the rate of transmission of intelligence on any one frequency is directly proportional to the band width used; (2) for a normal uniform energy distribution of noise over the frequency spectrum, the amount of noise energy received is directly proportional to the receiver band width; and (3) it has been operationally observed that present high-speed craft (including aircraft) can satisfactorily utilize navigational information furnished at a low rate of intelligence reception (e.g., 20-cycle modulation) by utilizing integration and average indication systems. Specifically, amplitude modulation may vary from CW through audio modulation to pulse modulation transmission. *CW transmission* provides no intelligence information except the direction of propagation of the wave; the latter information is extracted by the receiving station and is not a function of the type of modulation. The receiver band width required is extremely narrow, thereby reducing band-width noise to the minimum. Receiver-introduced modulation methods, as described on p. 22-42, may be used to discriminate further against noise and effectively increase the signal-to-noise ratio without affecting the receiver noise (which is determined by the input circuits). Large bearing errors may be experienced in pure CW systems at long range, where sky-wave transmission must be used. These bearing errors cannot be detected easily but may be reduced in certain cases by means of the receiver modulation methods mentioned. CW transmissions may be identified by their frequency or by means of very slow keying or on-off schedules which theoretically require a definite band width but can be reduced to a few cycles, effectively approaching zero.

AUDIO MODULATION can be taken to include modulation frequencies from a few cycles to several thousand cycles and does not impose prohibitive requirements on transmitter or receiver design. However, in view of the fact that the rate of intelligence transmission need not be high for most navigating systems, it would seem preferable to use band widths of the order of 10 to 100 cycles and utilize suitable receiving methods in order to reduce receiver band-width noise to the minimum. It has been operationally observed that modulated radio signals have a higher stability and accuracy performance beyond the ground-wave propagation range than do CW radio signals. This is apparently due to the operator's ability to distinguish single-path transmission from multiple-path transmission (either by aural or visual methods) when modulated signals are employed. With CW radio signals poorer performance results because it is practically impossible to distinguish the more stable single-path ray from the multiple-path rays with their larger instabilities and errors. Improved signal-to-noise ratios on desired signals can also be effected in audio modulated transmitting systems by employing receiver synchronizing procedures to match the transmitted modulation rate, thereby integrating or averaging out random noise. Furthermore, improvements in signal-to-noise ratios (e.g., bearing resolution and accuracy performance) can also be obtained in audio modulated systems by employing the same receiver-introduced-modulations technique discussed under CW systems and described on p. 22-42 under direction finders. Modulated systems also permit convenient means for positively effecting station identification either by keying, time scheduling, or variations of the modulation rates. It is important to note that modulation rates lend themselves to accurate synchronization between transmitters and receivers, not only for improving the signal-to-noise ratio but also for providing a highly effective means for efficiently employing the frequency spectrum. The spectrum can be conserved by permitting a number of stations to operate on a common carrier frequency with different known and stabilized modulation rates, allowing the receiver to select the desired station by synchronizing on the proper modulation. This procedure is already in use in various adaptations as in the Loran method of synchronized pulse modulation, where as many as 16 separate stations in a given area may share the common carrier frequency and are distinguishable at the receiver by the different synchronous repetition (i.e., modulation) rates. *Pulse modulation* systems require a greater band width, and, consequently, the inherent receiver noise is greater. However, methods of synchronous pulse gating are being developed which effectively reduce the noise and provide a higher sensitivity to pulse detection. The wide band required by one transmitter employing pulse modulation can be offset as in audio modulation by utilizing a number of precisely controlled repetition rates at the same carrier frequency (as mentioned above in the case of Loran), and by providing for differentiation between these various rates at the receiver. The greatest single advantage of pulse modulation is the very precise and highly convenient technique available for time synchronization, which permits effective time differentiation between the various transmission paths of the arriving wave. The improved signal-to-noise ratio advantage offered by modulation and pulse synchronization techniques constitutes another great advantage. However, pulsed techniques become more difficult to apply as the frequency decreases.

The efficiency of pulse synchronization techniques has been operationally proved in several instances, such as (1) in the India Theater during World War II where Loran operation was continuously maintained during the monsoon weather, when all communication links were erratic or inoperative; and (2) under conditions of severe precipitation static in Greenland when, again, Loran provided the only operative navigation link, whereas all communication circuits became unintelligible. The advantages of pulse techniques for maintaining continuous and reliable service must, therefore, be thoroughly investigated and included in any comprehensive analysis of the final choice of the operating parameters for a long-range world-wide navigation system.

Other modulations in addition to the various types of amplitude modulation exist, such as frequency modulation, phase modulation, and pulse-time modulation. Although no extensive study has been made of these for long-range navigational purposes it is well known that frequency and phase modulation possess considerable noise-reducing properties in addition to the modulation synchronization techniques previously described for improving signal-to-noise ratios. However, these advantages are partly offset by their usual wide-band transmission requirements and their general limitation to the higher frequencies (in order to achieve adequate depth of modulation) where line-of-sight range restrictions prevail. *Pulse-time modulation* (PTM) offers the advantage of increased intelligence on a pulse channel but decreases the number of simultaneous channels allowable at one frequency. It also makes the separation of single and multiple path transmissions more difficult and prevents use of the precise synchronization techniques necessary for effective improvements in signal-to-noise ratios. Lastly, hybrid or combination types of modulation and second and third derivative frequency modulation may play increasingly important roles as investigation continues into the techniques of such systems.

BIBLIOGRAPHY

1. Comdr. Benjamin Dutton, *Navigational Nautical Astronomy*. U. S. Naval Institute (1942).
2. *Aids to Marine Navigation*. U. S. Coast Guard (1940).
3. Marine Radiobeacons, *Engineering Instructions*, Chapter 41. U. S. Coast Guard (1944).
4. R. Keen, *Wireless Direction Finding*. Iliffe & Sons, London (1938).
5. M. K. Goldstein, Naval Research Laboratory, R-1896 (*HF/DF aboard the U.S.S. CORRY-DD463*); R-1938 (*Specifications for Minimum DF Site Requirements*); R-2229 (*Interpolation Charts for HF/DF Shipboard Calibrations*); R-2469 (*HF/DF Collector Locations on a PV-1 Navy Patrol Bomber*). Available through the U. S. Department of Commerce, Washington, D. C.
6. M. K. Goldstein, Naval Research Laboratory, Direction Finder patent applications, Serial Nos. 593,900 (Filed: 5/15/45, Navy Case 5280) and 443,899 (Filed: 5/5/42, Navy Case 3176); and Patent 2,405,203 issued 8/6/46.
7. Fundamentals of Radar, *Wireless World*, October 1945, pp. 299–302; November 1945, pp. 326–329; December 1945, pp. 363–365; January 1946, pp. 23–26; February 1946, pp. 55–56.
8. E. G. Schneider, Radar, *Proc. I.R.E.*, Vol. 34, No. 8, 528–578 (August 1946).
9. John H. DeWitt, Jr., Technical and Tactical Features of Radar, *J. Franklin Inst.*, February 1946, pp. 97–123.
10. *Principles of Radar*. M. I. T. Radar School (1944).
11. *Electronic Navigation Aids*. U. S. Coast Guard, Public Information Division (April 1946).
12. *Minimum Specification Brief Nos. 1, 2, and 3 Submitted as a Guide for Voluntary Use by Parties Interested in Navigational Radar*. U. S. Coast Guard (Nov. 14, 1945, revised Aug. 1, 1946).
13. J. A. Pierce, An Introduction to Loran, *Proc. I.R.E.*, Vol. 34, 216–234 (May 1946).
14. The Loran System, *Electronics*, Vol. 18 (November 1945); Vol. 18 (December 1945); Vol. 19 (March 1946).
15. John E. Clegg, Consol, Beacon Direction Finding Systems of High Accuracy, *Wireless World*, July 1946, pp. 233–235.
16. Report of Electronic Subdivision Advisory Group on Air Navigation, *Bulletin* 200, ATSC Eng. Div. (February 1946). Available through U. S. Army Air Forces, Wright Field, Dayton, Ohio.
17. A. Seeley, Shoran Precision Radar, *Electrical Engineering* (*Transactions*), April 1946, pp. 232–240.
18. M. K. Goldstein, Proximity Effect of Metallic Fences on Direction Finders, *Naval Research Laboratory Report R-1890*.
19. Spinning Loop Type Direction Finders proposed by M. K. Goldstein, 1942, see *Naval Research Laboratory Report R-2847*.
20. M. K. Goldstein, Facility and Installation Requirements for DAJ/DF Coastal Networks, *Naval Research Laboratory Report R-2079*.
21. Early Echo Ranging Sonar, *Electronics*, Vol. 19, 274–282 (July 1946); Echo Ranging Sonar, *Electronics*, Vol. 19, 88–93 (August 1946).
22. Lt. Comdr. D. H. Macmillian, RNR, *Precision Echo Sounding and Surveying*. Henry Hughes & Son, London (1939).
23. E. N. Dingley, Jr., A True Omnidirectional Radio Beacon, *Communications*, Vol. 20, 5, 6, and 35 (January 1940).
24. Papers on Detection of Infra-Red Radiations Given at Winter Meeting of The Optical Society of America held March 1946. Abstracted in *Journal of the Optical Society of America*, Vol. 36, No. 6 (June 1946).
25. G. C. Southworth, Microwave Radiation from the Sun, *J. Franklin Inst.*, Vol. 239, 285–297 (April 1945).
26. Frequency Power and Modulation for a Long-Range Radio Navigation System, *Electrical Communication*, Vol. 23, 144–158 (June 1946).

SECTION 23

MEDICAL APPLICATIONS OF ELECTRICITY

BY

CHARLES WEYL
S. REID WARREN, JR.

MEDICAL APPLICATIONS OF ELECTRICITY

By Charles Weyl and S. Reid Warren, Jr.

During the latter part of the eighteenth century Galvani and Volta discovered that the muscles of dead frogs' legs could be made to move by contact with metals. The descriptions of their experiments indicated that they were establishing a difference of electric potential between nerves and/or muscles of the frogs' legs and that this applied difference of potential caused a motion of the muscles. Since that time, many techniques have been devised for using electrical apparatus for studying physiological functions, for the diagnosis of disease, and for therapeusis. During the twentieth century the development of these medical applications of electricity has become accelerated and, particularly since 1925, new methods have been developed, tested, and applied to clinical practice. Although some of these techniques have become standardized by accurate experimental investigations, nevertheless, the biophysical explanations of many of the techniques are almost completely unknown. Although frequently the electrical apparatus and the techniques for their use are simple, an accurate knowledge of the results of the technique has been impeded by differences in individual patients, by lack of knowledge of the physiological factors involved, and by a lack of proper control of the apparatus and technique. Thus there exists a fertile field for investigation in which physicists and engineers are aiding physicians in problems of electromedical diagnosis and therapeusis.

With the accelerated development of the electromedical methods which were known before 1925, and also of the new methods which have been introduced since that time, there has been a great increase in the literature of the field. Furthermore, the applications of many of the newer devices and techniques to medicine are subsidiary to their applications in other scientific fields. This section is organized on the basis of these factors. Each article contains the following items: (1) a list of the fundamental components of the electrical apparatus required for a particular electromedical technique and an example of their combination into a practical, useful apparatus; (2) a brief description of the electromedical technique and examples of its use; (3) a list of references to papers and books in which the electromedical apparatus and technique are discussed in detail.

The literature of the medical applications of electricity uses terms quite different from those employed by physicists and engineers, and some of them are extremely confusing. For example, the term *electrotherapy* is sometimes used as a generic to represent nearly all the electromedical techniques except those involving x-rays; it also often represents solely those techniques described in the next article involving the conduction through parts of the body of direct, pulsating, and alternating currents.

ELECTROTHERAPY AND SHOCK THERAPY

1. APPARATUS

The apparatus employed for generating direct and alternating currents for electrotherapy has many forms, all of which are fundamentally of simple design from the engineering point of view. The requirements are as follows: (1) direct current, 0 to 80 volts, 0 to 50 ma, with a continuous resistance control; (2) alternating current at commercial frequencies from 0 to 50 volts rms, 0 to 25 ma; (3) a method for periodically surging or modulating the alternating currents mentioned in (2); (4) an induction coil operated by direct current with a mechanical interrupter; (5) a device for providing pulsating current at frequencies from 10 to 100 pulses per second.

Occasionally an attempt is made to construct an apparatus which will supply all these wave forms. Usually, however, the physician purchases different pieces of apparatus for each purpose. At present, the trend is toward the construction of apparatus which operates from an a-c source of 110 volts, 60 cycles. For example, the d-c generator may take the form of a small 1:1 ratio transformer with the secondary connected to a full-wave rectifier and filter system. Across the filter system is connected a potentiometer, and the electrodes are connected to two terminals of the potentiometer. The surging devices are generally motor-driven variable resistances. The apparatus used for these purposes has not been standardized and therefore it is not possible to give typical wiring diagrams.

The arrangements for protection of the patient from excessive electric shock are, in many installations, inadequate.

A number of accessory parts comprising, for example, specially designed electrodes and a timer for controlling the length of treatment are required.

The development of technique in electrotherapy is based entirely on empirical results frequently acquired from doubtful interpretations of meager data. Since remarkably little attention has been paid to the precise measurement of voltage and current, many of the accumulated data are of doubtful value.

Treatments usually have a duration of 1 to 15 min. In general the physician specifies his technique by a measurement of the current which flows through the patient, and the duration of the treatment.

2. ELECTROTHERAPY

It has been known for many years that passage of electric currents through the human body may cause important physical and physiochemical changes within the tissue. The following results form the basis of methods of diagnosis and treatment by the passage of electrical currents through the body: (1) the transfer of ions from outside the body into the treated part; (2) the transfer of ions from within the body to an electrode outside; (3) the production of heat within the living tissue; (4) the stimulation of nerve and muscle fibers.

To accomplish these purposes, direct currents, pulsating currents, and alternating currents are used. The following paragraphs describe the applications of each of these several forms of current.

ELECTROCHEMICAL CAUTERIZATION. A flow of direct current (called galvanic current in electrotherapy) through the body generally serves its therapeutic purpose by destroying diseased tissue. If platinum electrodes are applied to a diseased part which is near the surface of the body, and a direct current is made to flow between the electrodes, an acid tends to form by electrolytic action at the anode, and an alkali at the cathode. The destructive action of such chemicals is well known, and frequently it is possible to control the destruction of tissue by means of the electrical method for the production of these chemicals. Such a procedure is called electrochemical cauterization. Usually only one metallic electrode is placed in contact with the tissue. The other electrode, called the indifferent electrode, consists of a metal disk several inches in diameter held against a pad of cotton soaked in a solution of sodium chloride and placed against the skin of the patient. The transfer of ions from the salt solution apparently has negligible effect on the human body. Electrochemical cauterization is used for removing superfluous hair, warts, and small rodent ulcers, and also in the treatment of certain skin diseases.

IONTOPHORESIS. If one hand of the patient is placed in a liquid electrolyte, and this electrolyte is used as an electrode, it is possible by the application of direct current to transfer ions from the solution into the body. This process is called iontophoresis. The physiological results of treatment by this method have not been adequately explained. Some work has been done for the purpose of effecting local anesthesia by the transfer of ions from cocaine into the body at chosen places. Metallic ions have been introduced, and it is claimed that the recombinations that occur within the body provide small quantities of atomic metal or of metallic salts at the site of disease. Other chemicals such as acetylbetamethylcholine chloride and histamine have been used for the treatment of specific disorders.

CATAPHORESIS. If a semi-solid electrolytic gelatin is formed in the shape of a cylinder and two metal plates are placed at the ends of the cylinder, a direct current may be made to flow through the gelatin. After a few minutes have elapsed, there is an increase of water at the cathode and a decrease of water at the anode. This phenomenon is apparently caused by changes in osmotic pressure due in turn to the current flow. This procedure is sometimes used to remove undesired liquids from skin lesions. It is called cataphoresis.

MISCELLANEOUS. It has been found that muscles which are in a state of fatigue become considerably strengthened after treatment with direct current. It has also been claimed by many workers that the passage of current causes exhilaration of some patients, and, on the other hand, with a slightly different technical procedure, the same type of treatment may produce drowsiness. The causes of these effects are unknown.

The passage of small direct currents through the body causes no pain. However, a sudden change in the amplitude of the current produces muscular twitching and pain. Normally, d-c treatments require a source capable of adjustment from 0 to 80 volts. The current passing through the body may be from $^1/_2$ to 50 ma. Because of the muscular twitching described above, it is particularly important that the current be adjusted slowly to the desired value. Quantitatively, it is usual to increase the current linearly with time at the rate of 1 ma per min. The d-c resistance of the two arms and the trunk of the average adult is approximately 1400 ohms. This resistance is measured by suspending the hands in salt solutions, thereby overcoming the high resistance of dry skins.

The passage of alternating current of frequency less than a few thousand cycles per second involves a sensation of pain and twitching of the muscles. The intensity of muscular reaction and pain increases with the amplitude of the current and with the frequency up to an indefinite limiting value of several thousand cycles per second. When the frequency is raised above this value, the sensation decreases and finally disappears unless the amplitude of the current is sufficient to produce appreciable heat.

The steady application of alternating currents of a few milliamperes of commercial frequencies is extremely painful and dangerous. Therefore a device is required to produce a surging sinusoidal current which may be described in the following manner. At the beginning of the treatment the alternating current is zero amplitude. Over a period of the order of 2 sec it is increased to a maximum and decreased to zero. There follows a rest period of about 2 sec, and then the surge is repeated. Apparently the application of 60-cycle alternating currents causes contraction and extension of the muscles. Treatment by this method has been recommended for paralyzed muscles or muscles damaged by disease.

FARADIC AND MISCELLANEOUS WAVE FORMS. Various other types of alternating and pulsating currents have been recommended for therapeutic use. It will suffice to list a few of these: (1) the secondary current from an induction coil, called faradic current in electrotherapy; (2) currents produced by the periodic charge and discharge of condensers; (3) pulsating current of various wave forms produced by mechanical interrupters; (4) high-frequency brush discharge by means of a Tesla coil. Standardization of such methods will require a careful oscillographic analysis of the electrical parameters combined with a statistical analysis of the clinical results.

3. SHOCK THERAPY

In certain mental disorders it has been found that if a convulsive shock is produced in the patient by means of drugs, or by electric currents passing through the brain, improvement in the condition of the patient is often observed. The techniques of shock therapy are still based chiefly on empirical data. It is to be noted that the threshold at which electric shock of the brain will cause convulsion is relatively close to the higher lethal threshold. It is, therefore, necessary to control the shock precisely. Shock treatments are usually given in a series of from six to twelve treatments during periods of a few weeks.

It is found that, using electrodes placed approximately over the two temples, currents of the order of 1 amp, adequate to produce convulsion, are caused to flow by a 60-cycle alternating voltage of approximately 100 volts. Thus, the apparatus for shock therapy consists essentially of a transformer, the output of which can be continuously varied, and a timer capable of producing exposures from 0.1 sec to 0.6 sec in steps of 0.1 sec. Motor-driven timers, time-delay relays, and electronic timers have been used to produce the desired intervals of shock. In order to avoid fatal accidents, it is important in these devices to use circuits such that the failure of a component will result in a great decrease in the output voltage.

BIBLIOGRAPHY

Cumberbatch, E. P., *Essentials of Medical Electricity*. Mosby (1929).
Glasser, Otto, Editor, *Medical Physics*. Year Book Publications, Inc., Chicago, Ill. (1944). Physical therapy, low-frequency currents, pp. 1068–1073.
Kovacs, Richard, M. D., *Electro-therapy and the Elements of Light Therapy*. Lea and Febiger (1932).
Osbourne, S. L., and H. J. Holmquest, *Technic of Electro-therapy and Its Physical and Physiological Bases*. Charles C. Thomas, Springfield, Ill. (1944).
Electric Convulsion Therapy in Mental Disorders, *Psychiatric Quart.*, Vol. 14, 719 (1940).

DIATHERMY AND HIGH-FREQUENCY SURGERY

4. APPARATUS

Generators of high-frequency alternating current, from 750 kc per sec to 3000 Mc per sec, have been used to raise the temperature of parts of the human body in the techniques known as *diathermy* and *high-frequency surgery*.

There are two general types of diathermy apparatus. In the first type, a high-voltage transformer is connected to an oscillatory discharge circuit consisting of a spark gap, condensers, and an inductance. Loosely coupled to this oscillatory circuit is a secondary circuit to which the patient is connected. A high-frequency ammeter is connected in series with the patient to measure the magnitude of the treatment current. Figure 1 is a diagram of such a generator. More recently vacuum-tube oscillators (see Section 7) have

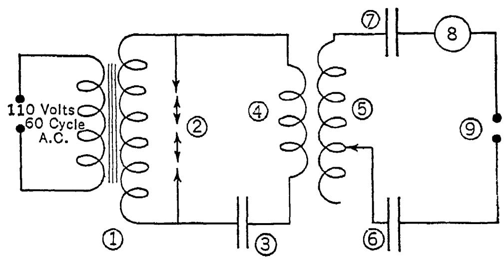


FIG. 1. Diathermy Apparatus Using Spark-gap High-frequency Generator

1, Step-up transformer; 2, spark gaps; 3, condenser, and 4, inductance forming oscillatory circuit which resonates at 500–1000 kc; 5, coil coupled to oscillatory circuit; 6, 7, condensers; 8, high-frequency a-c ammeter; 9, leads to electrodes.

been designed for this purpose. They are connected in a manner similar to that shown in Fig. 2. It will be noted that the two circuits are identical in all essential features with the two types of radio transmitters known respectively as spark and continuous-wave transmitters.

Various techniques are used for producing a high-frequency electromagnetic field within the part that is to be treated. Thus, the output of the oscillator is sometimes applied to two metallic plates between which the part of the body to be treated is placed. Alternatively the output of the oscillator may be fed to a cable which can be wrapped in the

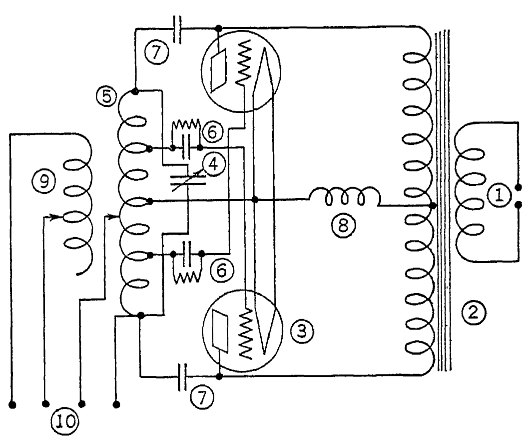


FIG. 2. Diathermy Apparatus Using Vacuum-tube Oscillator

1, 110-volt, 60-cycle a-c supply; 2, plate voltage transformer; 3, vacuum tube; 4, 5, oscillating circuit; 6, grid leak and grid condenser; 7, blocking condenser; 8, r-f choke; 9, inductively coupled output coil; 10, output leads to ammeter and patient. The cathodes of the vacuum tubes are heated by current from an additional secondary coil (not shown) of transformer 2.

form of a solenoid around a body part or shaped into a spiral coil (the technical term is pancake) which can then be fixed against one side of the part to be treated. The special electrodes required for surgery are described in article 6.

Apparatus using a magnetron operating at 2450 Mc is now available. Radiation is emitted by one of several parabolic reflectors 6 to 10 cm in diameter excited by a dipole. The radiation is thus confined to a relatively sharp beam.

Because of the very important danger of interference with communication facilities, the F.C.C. has prescribed the frequencies permitted for diathermy and the frequency stability that is required. The frequencies so assigned are: 13,560 kc ± 6.78 kc; 27,120 kc ± 160.00 kc; 40,680 kc ± 20.00 kc; 2450 Mc ± 50 Mc, available for experimental work.

5. DIATHERMY TECHNIQUE

If the frequency of the alternating current applied to the human body is indefinitely increased, an ill-defined point can be found (order of magnitude of 10,000 cycles per second) beyond which no sensation other than warmth is felt. The minimum amplitude of alternating current which produces pain by passage through the human body varies with the frequency of the alternating current and with the subject. Average values obtained by a number of workers are as follows:

Frequency, cycles per second		Average Tolerance, milliamperes
60	....	3– 8
10,000	tolerance current, increasing gradually to.......	25– 30
100,000	tolerance current, increasing quickly to.........	250–600

Heating effects, only, are noted above 100,000 cycles per second. As a result it is possible to increase the current flowing through the body to values as high as 5 amp. This causes a sensible dissipation of heat within the tissue. For this purpose frequencies above 750 kc per sec are used. It has been discovered that heat produced internally is effectual in the treatment of certain diseases. The name *diathermy* has been given to the treatments characterized by the internal heating effect of high-frequency alternating currents. Electrodes several square centimeters in area are used in order that the current may be distributed approximately uniformly over a large area. Other machines in which the patient is not in contact with electrodes, but is placed in a strong high-frequency electromagnetic field (between "condenser" electrodes), have been constructed for the purpose of heating the entire body or, more commonly, large portions of the body. It is possible to raise the body temperature as much as 3 or 4 deg C by this means. The method has proved useful in the treatment of syphilis, pleurisy, neuritis, and other diseases. The currents normally used range from 0.5 to 5.0 amp.

6. HIGH-FREQUENCY SURGERY

If one large electrode is applied to a part of the body and a second electrode consisting of a large platinum needle is brought in contact with any other part of the body, the resultant current density is so high at the point of application of the needle that the tissue may be completely destroyed by heat. This method, which is called electrosurgery, is a relatively recent development. The small blood vessels severed by means of the high-frequency knife are normally sealed by the heating action, and bleeding is therefore considerably reduced. Basically, there is no difference between the equipment used for diathermy and for high-frequency surgery.

BIBLIOGRAPHY

Cumberbatch, E. P., *Diathermy*. London (1929).
Kovacs, Richard, M. D., *Electro-Therapy and Light Therapy*, Lea and Febiger (1932).
Osbourne, S. L., and H. J. Holmquest, *Technic of Electro-Therapy, and Its Physical and Physiological Bases*. Charles C. Thomas, Springfield, Ill., pp. 315–744 (1944).

THE MEDICAL USES OF ULTRAVIOLET AND INFRARED RADIATIONS

7. APPARATUS

Apparatus for the production of ultraviolet radiation includes the carbon-arc generator, the quartz-enclosed mercury-arc generator, and the mercury glow-discharge lamp enclosed in quartz or Corex, a glass which transmits a useful amount of ultraviolet radiation. One of the chief objections to the carbon-arc generator is its production of relatively high-intensity visible and infrared radiations. In addition to the above-mentioned lamps, which have relatively high outputs, several special lamps have been developed for home use under medical supervision.

One of these—the Mazda RS lamp—is rated at 110 volts 275 watts. At a distance of 24 in. this lamp will produce a mild reddening (erythema) of the skin in about 5 min. The lamp is strong in those radiations that produce sunburn (2800 to 3200 angstrom units). There is also a relatively strong infrared beam produced chiefly by a tungsten filament in an atmosphere of argon and nitrogen.

The outer envelope of the Mazda RS lamp is made of special glass that transmits infrared and visible radiation, and ultraviolet radiation of wavelengths greater than 2800 A. A reflecting surface is deposited on the interior surface of this envelope; the radiation is emitted through the circular (12-cm diameter) end of the lamp opposite the screwplug. A tungsten filament, which operates as a series ballast after the initial preheating period, is mounted in the outer envelope; the space is filled with argon and nitrogen.

Within the larger outer envelope, there is a quartz capsule containing an oxide-coated filament and a second, initially cold, electrode; this capsule contains mercury.

When the lamp is initially connected to 110-volt 60-cycle alternating current, the outer (tungsten) filament and the inner (oxide-coated) filament are connected in series. After approximately 30 sec, when the oxide-coated filament acquires a sufficiently high temperature for copious emission of electrons, a thermal delay switch within the tube operates to connect the initially cold electrode of the quartz capsule to the terminal of the supply opposite from the terminal connected, through the tungsten filament, to the oxide-coated filament and disconnects the oxide-coated filament. The mercury arc then strikes, becoming stable as the initially cold electrode becomes heated by bombardment. The arc in the mercury vapor at a pressure of approximately 1.1 atmospheres then operates in series with the tungsten filament; the filament and arc emit ultraviolet, visible, and infrared radiation. The manufacturer recommends a 1-min. "warm-up" period before use.

Glow-discharge mercury tubes operating at voltages as high as 5000 volts and currents of the order of 15 ma produce radiation almost entirely of the emission frequency line of mercury having a wavelength of 2537 A. This radiation is strongly bactericidal. Since it also requires low-power input, this type of lamp is often used to irradiate the area in operating rooms and in other places where air-borne bacteria are to be minimized. Experiments using these devices in schoolrooms are being conducted.

Since the mercury-arc lamp enclosed in quartz produces a relatively high-intensity radiation at wavelengths from 2483 to 4047 A, and since this apparatus is relatively stable in operation after a few minutes of operation, it is now the most common source for medical purposes. The total power input to such a device is from 250 to 400 watts, and its output of ultraviolet radiation is of the order of 225 microwatts per square centimeter at a distance of 1 meter without a reflector.

Special photoelectric cells connected to integrating circuits and counters have been devised to measure the outputs of ultraviolet generators throughout various bands of wavelengths. However, such standardizing measurements have not been very often applied to clinical use.

Infrared radiation is produced chiefly by electrically heated wire operating at temperatures from 800 to 1600 deg C or by means of incandescent tungsten lamps with or without filters.

8. THERAPEUTIC USE OF ULTRAVIOLET RADIATION

It has been known for many years that sunlight has a definitely beneficial effect upon the human body. Since 1800 the spectra of light emitted by the sun and by electric arcs have been analyzed and more attention has been directed to the specific photochemical effects of the various parts of the spectrum. More or less arbitrarily, ultraviolet radiation has been defined as radiation with a range of wavelengths from approximately 4000 to about 40 A (1 angstrom unit = 10^{-8} cm). The solar spectrum extends down to approximately 2900 A but under average atmospheric conditions at sea level there is negligible radiation below 3000 A. It has been shown that most of the ultraviolet radiation incident upon the human body from the sun or from artificial sources is absorbed in the surface tissues. The penetration is 0.1 mm or less.

There has been much speculation as to the biological processes produced by ultraviolet radiation, which has so far resulted largely in contradictions and disagreements among authorities on the subject.

The first noticeable effects of ultraviolet radiation are erythema, or reddening of the skin, usually followed by pigmentation. Each individual has a different tolerance to ultraviolet radiation. What may cause a pronounced erythema in one may produce no effect upon another. It has been discovered that certain non-soluble fats, particularly ergosterol, form vitamin D when irradiated by ultraviolet rays. This substance is present in the human body, and normal exposure to sunlight is one method for supplying the definite need of the body for vitamin D. Although there is no question as to the beneficial effects of sunlight and artificial sources of ultraviolet radiation when exposures are carefully controlled, enthusiasm on the part both of physicians and of laymen has led to dangerous overexposures to both natural and artificial sources of this form of energy. It has not been proved that exposure to direct sunlight is essential to the physical well-being of the normal healthy human.

Treatment by ultraviolet radiation is useful in cases of rickets, high blood pressure, and some skin diseases. It has been shown definitely that ultraviolet radiation produces directly and indirectly a substantial rise in the amount of calcium and phosphorus in the blood. The treatment is in general use for rickets in infants and has been remarkably successful. Certain experimenters believe that the effect of ultraviolet rays on the nervous system is stimulating. A number of psychiatrists have recommended the treatment for neurasthenia and some of the psychoses.

After Pasteur's discovery that small living organisms are the cause of many diseases and also of fermentation, experiments were performed to find the effects of various types of radiation upon these organisms. It was found that ultraviolet rays were potent in bactericidal action, particularly in the 2600 A region.

Different parts of the body vary considerably in sensitivity. For example, twenty to thirty times as much ultraviolet radiation is required to produce an erythema of the soles of the feet than of the face, which is the most sensitive part. Even small doses which penetrate to the eyeball can cause serious damage; therefore all patients receiving ultraviolet treatment are required to wear goggles made of glass which absorbs practically all the ultraviolet radiation. Practically no ultraviolet radiation is transmitted through window glass and very little through smoky atmosphere. Therefore sunlight may be used as a source of therapeutic ultraviolet rays to the best advantage in special regions—for example, in high mountains where there are generally very few clouds and practically no dust or smoke. Since the ultraviolet content of sunlight varies with latitude, season, time of day, and atmospheric conditions (such as ozone, water vapor, dust, and smoke) it is important to measure the intensity in the region of 2900–3200 A during treatment.

9. THERAPEUTIC USE OF INFRARED RADIATION

Infrared radiation is used for surface heating, either of a small circumscribed area or for the surface of an entire leg or arm. One of the chief effects is the production of increased blood flow near the surface, and this may, in turn, cause more deep-seated changes.

It is possible also to change the temperature of the entire body. Thus the body may be placed in a heat-insulated housing inside of which incandescent lamps or other heating elements raise the temperature of the body and of the area surrounding it. Since there is no way for this heat to be entirely radiated, the temperature of the body is caused to rise. Such whole-body heating may also be accomplished by means of diathermy.

BIBLIOGRAPHY

Bernhard, O., *Light Treatment in Surgery.* Arnold, London (1923).
Duggar, B. A., *Biological Effects of Radiation,* Vol. 1. McGraw-Hill (1936).
Finsen, N. R., *Phototherapy.* Arnold, London (1901).
Glasser, Otto, Editor, *Medical Physics.* Year Book Publications, Inc., Chicago, Ill. (1944). Physical therapy—heat and cold, pp. 1043–1054; light, pp. 1054–1068; radiation, sources of ultraviolet and infrared, pp. 1157–1163.
Laurens, Henry, *The Physiological Effects of Radiant Energy.* Chemical Catalog Co. (1923).
Rollier, A., *Heliotherapy.* Frowde, Hodder and Stoughton, London (1923).
Osbourne, S. L., and H. J. Holmquest, *Technic of Electro-therapy and Its Physical and Physiological Bases.* Charles C. Thomas, Springfield, Ill. (1944). Ultraviolet radiation, pp. 235–313; infrared radiation, pp. 194–234.

ELECTROCARDIOGRAPHY AND ELECTROENCEPHALOGRAPHY

10. APPARATUS

Muscular contractions and the functioning of nerves are accompanied by measurable differences of electric potential at the surface of the body. Thus the time-varying differences of potential between the right arm and the left arm, the right arm and the left leg, the left arm and the left leg are used by cardiologists to interpret the action of the heart; these differences of potential are called respectively the potentials of Lead I, Lead II, and Lead III. The measured differences of potential are of the order of a few millivolts. An electrocardiogram for Lead II for a normal patient and a heart sound record made simultaneously are shown in Fig. 1. For the electrocardiogram, a vertical deflection of one small division corresponds to a difference of potential of 0.2 mv.

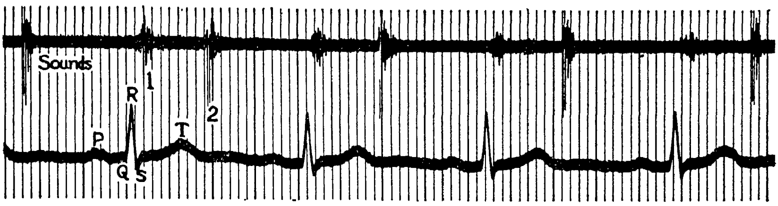


Fig. 1. Electrocardiogram

Lower.—Electrocardiogram, Lead II, normal patient. The letters P, Q, R, S, and T are standard symbols for designating the waves thus marked.
Upper.—Heart sound record made simultaneously with the electrocardiogram above.

Electrocardiographic apparatus requires the following component parts: (1) a device for transforming the electromotive force generated by the heart into the motion of a light beam or of a shadow; (2) a camera with moving film or paper to record the deflections of the light beam or shadow; (3) devices for calibrating the abscissa (time) and the ordinate (electromotive force) on the electrocardiogram.

Einthoven developed a system similar to that shown in Fig. 2. A silvered quartz string with variable tension is held perpendicular to an intense constant magnetic field produced by an electromagnet. A calibrated external source of electromotive force is used as standardizer. A variable resistance and associated battery are used to compensate for skin currents (described below under technique). A specially adapted camera, with a tuning fork or other timing device, is used to record the deflection of the string produced by the heart action.

It has been shown that the Einthoven string galvanometer produces distortion due to the change in its response with frequency; furthermore, as is described below, it requires, for an adjustment of sensitivity, a change in the string tension which produces further distortion. Therefore, efforts were made to use galvanometers with more uniform frequency response. It was found that the use of carefully constructed thermionic amplifiers in association with these new galvanometers provided instruments of great flexibility and stability. The cardiographic paper or film can be driven by a synchronous motor, so that the time abscissas are the same for all cardiograms.

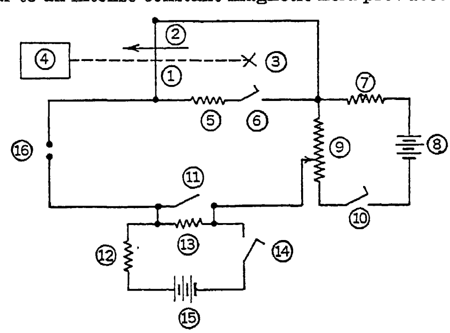


Fig. 2. Outline Diagram of Einthoven String Electrocardiograph

1, galvanometer conducting string of silvered quartz, suspended with controllable tension between the poles of an electromagnet (not shown) which produces a magnetic field in the direction of the arrow 2; 3, light source to obtain, with system of lenses not shown, an image of the string upon the moving photographic film in the camera 4; 5, 6, a shunt, of about one-tenth the resistance of the string, and a switch; 7, 8, 9, 10, the compensator, a series circuit which introduces an adjustable electromotive force in series with the string to compensate for skin current; 11, 12, 13, 14, 15, a circuit for introducing a calibrating emf of 1 mv in series with the string; 16, leads to the patient.

It is most convenient to describe the technique of electrocardiography by reference to the Einthoven instrument shown in Fig. 2. A resistance 5 is connected in parallel with the string, reducing its sensitivity by a ratio of 10 to 1. One area on each arm of the patient is rubbed with a contact paste and a metal electrode is applied. This corresponds to the designation described above as Lead I. A switch is closed connecting the patient to the electrocardiograph. The galvanometer deflects, owing to what is called skin current. The slide of 9 is then moved until the galvanometer string is returned to its zero position. This operation balances out the skin current, which is constant and plays no part in the interpretation of the electrocardiogram. After the compensator is adjusted, the shunt 5 across the galvanometer string is disconnected. This increases the amplitude of swing of the fiber. The standardizing circuit is then connected by closing switch 14 and opening switch 11. The circuit 12, 13, 15 is designed to produce a potential drop of 1 mv across the resistance 13. This throws the image of the string across the screen. An adjustment of the mechanical tension of the string is made so that the application of the 1-mv standardizing voltage produces a deflection of exactly 1 cm. The standardizing voltage is then removed by opening switch 14 and closing switch 11, and the electrocardiogram for Lead I is taken by means of a paper moving at constant speed past the beam of light through the galvanometer. For each centimeter deflection of this record an electrocardiographic impulse of 1 mv is necessary. The same procedure is followed for Leads II and III. The moving photographic paper passes before a glass screen upon which lines 1 mm apart are ruled in the direction of motion of the paper, while a synchronous motor turns a bladed wheel through the light beam at right angles to the direction of motion of the paper, producing a series of time-marking lines spaced at 0.04-sec intervals.

The procedure described can be accomplished quickly in practice. The entire operation is illustrated in Fig. 3. This description applies to the older type of electrocardiograph, from which the underlying principles can be clearly understood. The routine use of a modern electrocardiograph is characterized by practical simplicity, although a description of this procedure would not illustrate so clearly the fundamental characteristics.

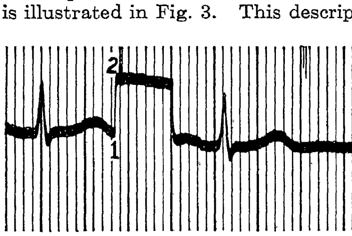


FIG. 3. Calibration of Electrocardiograph

The calibration record which is made at the end of the electrocardiogram of each patient. At (1) the standardizing voltage of 1 mv is connected in series with the galvanometer, causing a deflection of 1 cm to (2). The electrocardiogram is of Lead II, normal patient.

The modern electrocardiograph may be battery operated or it may operate from a 110-volt 60-cycle source. It is usually a self-contained portable device, and the records may be produced on film or photographic paper or—more recently—may be directly recorded on paper that may be examined as soon as it is produced. These devices are equipped with stabilized audio-frequency amplifiers feeding into specially designed galvanometers. The frequency characteristic of these electrocardiographs results in the reproduction, at nearly constant levels, of frequencies from 0.5 to 50 or 100 cycles per second. Because of the susceptibility of these high-impedance amplifiers to hum pickup, special hum-bucking input circuits are incorporated.

An electroencephalograph is used to record time-varying changes of potential between pairs of electrodes in contact with various parts of the scalp; the magnitudes of these differences of potential are from a few microvolts to approximately 100 mv. The component frequencies of the electroencephalographic signals are less than 1000 cycles per second. It follows, therefore, that an electroencephalograph can be a device similar to an electrocardiograph except that (1) the amplification must be about 100 times greater in the electroencephalograph, and (2) it is considered essential to provide fourteen-channel input (and two- to six-channel output) for the electroencephalograph instead of four-channel input which is used in the electrocardiograph.

11. TECHNIQUES

Since the middle of the nineteenth century it has been known that an electromotive force is generated within the heart during the period of contraction of this muscular organ. The precise causes of this effect are unknown. Nevertheless, the methods for measuring the variations of this electromotive force with time have been carefully standardized, and statistical records have been made for many years. By associating these records with case histories, a technique has been evolved for diagnosing certain diseases of the heart which is, in many instances, remarkably valuable. The development of electrocardiography has shown the benefits of carefully standardized scientific methods more than most other special fields of medicine in which electrical equipment has been used.

When the so-called "brain waves" are recorded by means of an electroencephalograph the wave forms are found to vary with age, with the somatic state of the individual, with the state of mental health of the individual, and with many other factors. Thus the development of this field of investigation has been based upon the recording of many electroencephalograms, their analysis and intercomparison, and finally attempts at diagnosis based upon these analyses. The electroencephalograph now finds clinical use in the corroborative diagnosis of epilepsy and, in many cases, in the localization of brain tumors.

BIBLIOGRAPHY

Ashman, Richard, and Edgar Hull, *Essentials of Electrocardiography*, The Macmillan Co., New York (1945).
Berger, H., *Archives of Psychiatry*, Vol. 87, 527 (1929); Vol. 100, 301 (1933).
Glasser, Otto, Editor, *Medical Physics*. The Year Book Publications, Inc., Chicago, Ill. (1944). Electrocardiography, pp. 352–360; electroencephalography, pp. 361–371.
Jasper, H. H., and L. Carmichael, *Science*, Vol. 81 (January 1935).
Likoff, W. B., M. B. Rappapart, and S. A. Levine, Continuous Recording Electrocardiography, *Am. Heart J.*, Vol. 28, 98–114 (July 1944).
Traugott, Paul, Electroencephalograph Design, *Electronics*, Vol. 16, 132–144 (August 1943).

ELECTROACOUSTIC DEVICES

12. AIDS TO THE DEAF

Those whose hearing is deficient require aids designed upon the basis of a quantitative study of their relative deafness. For this study a device called the audiometer is used; this instrument consists essentially of a vacuum-tube oscillator capable of producing practically pure tones accurately controllable as to pitch and intensity in head receivers. The pitch proceeds by octaves from 32 double vibrations per second to 16,384, with an intensity range of 10^9. Charts called audiograms have been made, showing the thresholds in sensation units, of hearing and of pain, for thousands of subjects with normal hearing. By comparing the audiograms, both the type and degree of deafness may be estimated. This constitutes a valuable aid in diagnosis and also is a useful adjunct for the prescription of hearing aids.

Many who are deaf learn to read lips. Care must be taken, therefore, to eliminate this source of error from tests of hearing aids, even though lip-reading may become a part of the method of understanding speech after the hearing aid is adopted. Some who have been deaf for a period of years require a considerable amount of time to learn to understand sounds heard with the aid of an electric device.

Hearing aids usually consist of a small microphone, a battery, and a special light-weight electromagnetic or crystal receiver. The system is designed to produce at the ear, as nearly as possible, an amplified replica of the sounds incident upon the microphone. If such a simple device does not amplify sufficiently, an amplifier using vacuum tubes may be added.

In recent years, the development of so-called "miniature" and "hearing-aid" tubes and of small batteries and printed circuits has considerably reduced the size and weight of such equipment.

More elaborate devices have been made for those whose hearing is especially deficient, in some instances the apparatus being designed to suit the individual requirements as determined from studies of the audiograms and the personal characteristics of the subject. Among the more elaborate examples of this type of hearing aid is the two-channel system in which two high-fidelity, velocity microphones are connected individually to two high-quality amplifiers, equipped with filters especially designed to compensate for the auditory deficiencies of the subject. The outputs of these two systems are then connected to individual headphones so that each ear of the subject has an entirely separate channel, separately connected, and affording corrected binaural (two-ear) hearing.

13. THE STETHOPHONE

About 1924 a device for picking up human heart sounds by means of a microphone and amplifying these sounds was developed and given the name "stethophone." The original purpose of the device was to make it possible for students in the amphitheater of the heart clinic to listen simultaneously with the demonstrator to the heart sounds of the patient under examination. The device had the additional advantage over the ordinary stethoscope of producing louder sounds in the listener's ear. It was further shown that by the insertion of electrical filters it was possible to eliminate portions of the audible frequency spectrum in order to concentrate upon certain specific sounds. This means of selective listening proved an aid in the diagnosis of murmers and other cardiac abnormalities. More recently this device has been used by physicians who are somewhat hard of hearing.

RECORDING OF HEART SOUNDS. Heart specialists have also shown that it is of value to record the heart sounds on a moving strip of paper or film simultaneously with an electrocardiographic record. This dual record has proved an aid in the diagnosis of certain cardiac disorders, and also for purposes of research into the mechanism of the heart.

BIBLIOGRAPHY

Cabot, R. C., and H. F. Dodge, Frequency Characteristics of Heart and Lung Sounds, *J. Am. Medical Assoc.*, Vol. 84, 24 (June 1945).
Fletcher, Harvey, *Speech and Hearing*. Van Nostrand (1929).
Frederick, H. A., and H. F. Dodge, The Stethophone, *Bell Sys. Tech. J.*, Vol. 3, 531–549 (October 1924).
Gamble, C. J., and D. E. Replogle, A Multiple Electrical Stethoscope for Testing, *J. Am. Medical Assoc.*, Vol. 82, 388 (February 1924).
Margolis, A., *Archives of Internal Medicine*, Vol. 46, 1048 (December 1930).
Rappapart, M. B., and H. B. Sprague, Graphic Registration of Heart Sounds, *Am. Heart J.*, Vol. 23, 591–623 (May 1942).
Williams, H. B., and H. F. Dodge, Analysis of Heart Sounds, *Archives of Internal Medicine*, Vol. 38, 685–693 (December 1926).

ROENTGEN THERAPY

14. PURPOSE AND GENERAL TECHNICAL REQUIREMENTS

A short time after the announcement by Roentgen of the discovery of x-rays in 1895, a number of physicians began to use the new radiation for the treatment of certain diseases. It has been only since 1920, however, that important detailed data have been recorded and correlated. It has been definitely determined that the absorption of x-rays by living tissue can cause the destruction of that tissue. Destruction of a given type of tissue is dependent upon the dose rate (in roentgens per minute; see below), the duration and frequency of treatments, the area of the irradiated skin surface (portal), the total dose, and probably other factors as well. The destructive action is radically different with different kinds of tissue. The reproductive cells of the human body are most sensitive, and the bones are least affected. The treatment of malignant tumors, such as cancer, is based upon the fact that, to a certain extent, it is possible to cause destruction of the diseased tissue without permanently damaging normal surrounding tissue.

X-rays generated by means of a hot-cathode tube have continuous spectra characterized by a minimum wavelength dependent upon the maximum x-ray tube voltage. The shorter the wavelength, the more penetrating is the radiation (see also Section 4, x-ray tubes). For these reasons the kind of x-ray equipment chosen for therapeutic use depends upon the site of the diseased portion which it is intended to treat. If, for example, it is desired to treat the skin, it may prove desirable to use radiation of a wavelength of the order of 1 A. Such radiations are sometimes called Grenz rays. The most generally used x-ray therapy apparatus is operated at x-ray tube voltages of 50 to 400 kvp (peak kilovolts). A filter consisting of a few millimeters of aluminum, a few tenths of a millimeter of copper, or a combination of the two, is inserted between the x-ray tube and the patient. This filter absorbs a large part of the low-frequency radiation which would otherwise be absorbed by the skin. The filtered x-ray energy penetrates to the site of disease, and a reasonable proportion is absorbed and helps to produce the desired effect. Treatments by x-ray therapy have been apparently beneficial in some cases, particularly if the diseased part is properly diagnosed early in its development.

The gamma rays of radium have been valuable in treating cancer. It has therefore been assumed that x-ray apparatus capable of generating radiation comparable in frequency to gamma rays might prove useful. Experiments in this direction have led to the construction of tubes and apparatus capable of operating at 600 kvp to 2 Mev.

APPARATUS. The high-voltage generator (50–400 kvp) required for exciting an x-ray therapy tube is usually constructed with one of the following typical circuits as a basis (see also Section 7, Power Supply): (1) half-wave thermionic rectification; (2) special half-wave rectified voltage-doubling circuits with condensers (see Fig. 1); (3) full-wave thermionic rectification with condensers (nearly constant potential).

Two kinds of apparatus have been devised and put into relatively common use for operation at 1 and 2 Mev.

In one of these devices a van de Graaff generator and an x-ray tube are assembled in a steel tank into which air is introduced at a pressure of several atmospheres, the spark-over potential gradient of gases at high pressures being considerably greater than at normal atmospheric pressure. The van de Graaff generator consists of a continuous belt of insulating material mounted on two pulleys, one of which is at ground potential and is driven by a motor. Near the bottom pulley electrodes are mounted in front of and in back of the belt. A difference of potential is applied to these electrodes so that negative charges (electrons) are deposited on the belt. Thus the moving belt effectively carries a negative charge toward the top pulley. Surrounding this upper pulley there is mounted a hollow metal electrode to which a brush near the belt is connected. Electrons from the belt are conducted to the exterior surface of the hollow electrode, which thus acquires a negative charge. The cathode of the x-ray tube is connected to this upper electrode of the generator; the anode of the tube is grounded. In order to maintain uniform potential gradient throughout the length of the tube, cylindrical accelerating anodes are sealed into the glass column of the tube and connected to taps on a resistor which in turn is connected in parallel with the generator. The potential of the upper electrode with respect to ground increases at the beginning of operation until the sum of the resistor current, the x-ray tube current, and the leakage currents is equal to the rate at which charge is carried up on the belt. Thus, the tube operates at constant potential at voltages from 1 to 2 Mev, and currents of the order of a few tenths milliampere are obtained.

The second type of equipment used for voltages above 1 Mev employs the same kind of x-ray tube as the device described above but a different kind of high-voltage generator. The generator consists of an auto transformer (air core) operating at 180 cycles per second.

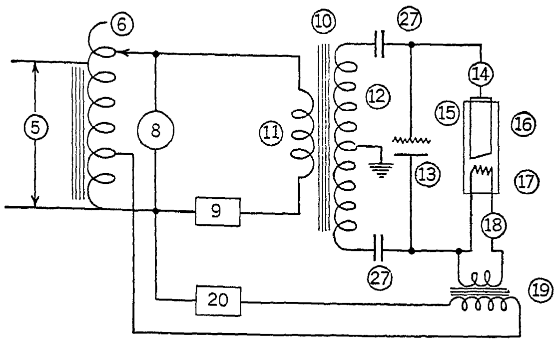


FIG. 1. X-ray Therapy Apparatus, with Half-wave, Voltage-doubling Circuit

5, a-c leads; 6, auto-transformer; 8, a-c voltmeter; 9, exposure timer; 10, high-tension transformer; 13, thermionic valve; 14, d-c milliammeter; 15, x-ray tube of which 16 is the anode, 17, the cathode; 18, cathode a-c ammeter; 19, the cathode heating transformer; 20, the cathode current regulator; and 27, the condenser.

The primary consists of a few turns of wire near the anode end of the tube wound coaxially with the axis of the tube. The secondary consists of a series of coils (each connected to an accelerating anode in the tube) mounted one above the other along the tube so that the top coil is approximately at the level of the cathode of the tube. This apparatus is enclosed in a steel tank into which "Freon" gas is pumped at a pressure of approximately 3 atmospheres.

15. TECHNIQUE

The x-ray tube mountings are arranged so that the patient may recline in a comfortable position during treatment. A lead-lined cone protects the patient and operator from scattered radiation. The radiation coming through the filter upon the patient is thus confined to an area of 10–400 sq cm called a *portal.* Investigations have been made to determine how much radiation is absorbed at various depths within the tissue. Sometimes in order to get the desired absorption within a deeply seated tumor it is necessary to turn the patient and to give several exposures through different portals, the central x-ray beam passing, in each case, through the tumor. This method of *cross-firing* prevents the absorption in any particular skin area from exceeding a tolerable dose. The distance from the tube to the patient is generally 50 to 100 cm. Tubes operate at currents of 5 to 30 ma, and the time of exposure may be from 5 to 45 min.

Just as with ultraviolet treatment, the individual tolerance must be investigated to prevent x-ray burn. In general a large factor of safety is allowed to prevent such a possibility.

The unit of x-ray dosage is called the roentgen. It is defined in terms of measurement by means of an ionization chamber. The roentgen is the quantity of x- or gamma-radiation such that the associated corpuscular emission per 0.001293 gram of air produces (in air) ions carrying 1 esu of quantity of electricity of either sign. It is assumed and apparently justified by empirical results that this measurement parallels the biologic effect of x-rays. The unit is an international standard. In some laboratories an ionization chamber is connected at all times; and in a few, the chamber operates an electric counter which integrates the total dose and turns off the power supply to the x-ray tube after the desired exposure. In other laboratories the x-ray machine is calibrated by means of a sphere gap and a milliammeter which measure respectively the peak kilovoltage supply to the x-ray tube and the average current through the x-ray tube. This calibration in turn is referred to the results of ionization-chamber measurements made several times each year.

BIBLIOGRAPHY

Charlton, E. E., W. F. Westendorp, G. Hotaling, and L. E. Dempster, New Million-volt X-ray Outfit, *J. Applied Phys.*, Vol. 10, 6, 374 (1939).

Duggar, B. A., *Biological Effects of Radiation.* McGraw-Hill (1936).

Glasser, Otto, Editor, *Science of Radiology.* Charles C. Thomas, Springfield, Ill. (1933).

Glasser, Otto, L. S. Taylor, Edith Quimby, and J. L. Weatherwax, *The Physics of Radiology.* Hoeber (1944).

Mayneord, W. B., *The Physics of X-ray Therapy.* Churchill, London (1929).

Robertson, J. K., *Radiology Physics.* Van Nostrand (1941).

van de Graaff, R. J., and J. G. Trump, Design of a Million-volt X-ray Generator for Cancer Treatment and Research, *J. Applied Phys.*, Vol. 8, 9, 602 (1937).

Weyl, Charles, S. R. Warren, Jr., and D. B. O'Neill, *Radiologic Physics.* Charles C. Thomas, Springfield, Ill. (1941).

ROENTGENOGRAPHY AND ROENTGENOSCOPY

16. PURPOSE, GENERAL TECHNICAL REQUIREMENTS, AND TECHNIQUE

In the first group of experiments performed by Roentgen the following x-ray phenomena were observed.

1. X-rays, incident upon photographically sensitive emulsions, produced a latent image similarly to visible light. Development of the emulsion produced a darkening of the film or plate throughout the area traversed by x-rays.

2. X-rays, incident upon barium platino cyanide, produced a visible (fluorescent) radiation.

3. X-ray absorption was greater for a given thickness of absorbing material in materials of high density than in materials of low density. For a given absorbing material x-ray absorption was found to be greater as the thickness increased.

These three facts form the bases of modern roentgenography.

The physically measurable characteristics of a roentgenogram of a part of the human body which are most important for medical diagnostic purposes are: (1) roentgenographic density; (2) roentgenographic contrast; (3) roentgenographic sharpness. Roentgenographic density and photographic density are identically defined as the logarithm to the base 10 of the ratio of light incident upon a particular area of film to the intensity of light transmitted through this area. Density is a function of x-ray intensity, therefore of x-ray tube voltage, distance of the plate from the x-ray tube, current through the x-ray tube, time of exposure, and of several less important factors, and also of the type of photographic material and of the method of development and the fixing and drying of this material. Density is, of course, a function of the physical characteristics of the object which is roentgenographed. Roentgenographic density may be measured with a polarization photometer or a photoelectric densitometer, or, for rougher approximation, densities may be compared by eye. This last method is most unsatisfactory. Roentgenographic contrast is the difference between the two densities of two areas of the roentgenogram and is therefore a function of the same variables as roentgenographic density. If it is desired to perceive the difference produced roentgenographically by tissue of nearly similar x-ray absorption characteristics in the object roentgenographed, then it is important to control the technique of roentgenography so as to make this difference clearly visible to the eye. Roentgenographic sharpness is the ability of a particular roentgenographic equipment to reproduce precisely borderlines between contiguous but definitely different densities. Information deduced from the physical measurements of these characteristics is influenced greatly by physiological and psychological factors associated with the viewing of roentgenograms, which factors have not as yet been completely investigated.

In order to roentgenograph any particular part of the human body a tube having a very small focal spot is necessary. It is also advisable to remove the film and object from the tube as far as possible in order to decrease distortion due to magnification of those parts of the object not in contact with the film. Since the focal spot of the x-ray tube is not a point there will be some loss in roentgenographic sharpness due to its finite size. If the object to be roentgenographed may be kept stationary for a moderately long period of exposure, the focal spot may be made correspondingly small. This method is employed for roentgenography of teeth and bones. If the part to be roentgenographed is continuously in motion a relatively short exposure is necessary, in order to arrest this motion sufficiently. Therefore, to roentgenograph parts such as the human chest the x-ray tube focal spot must be made correspondingly larger in order to dissipate rapidly the energy necessary for short exposure time. For roentgenography of the chest and heart, exposures of $^1/_{30}$ to $^1/_5$ sec are used. For roentgenography of other parts of the body the exposure time may be from $^1/_2$ to 20 sec.

To control the contrast in the roentgenograms the voltage of the x-ray tube is varied. The x-rays produced by high voltages are more penetrating than those produced by lower voltages. The voltage of the x-ray tube is generally measured by means of a sphere gap and is calibrated, for the particular current used, against the primary voltmeter reading. The x-ray-tube current for exposures longer than 3 sec is measured by a d-c d'Arsonval milliammeter. For very short exposures a ballistic milliampere-second meter is used to measure the total quantity of electricity passing through the x-ray tube. The timer for short exposures is arranged to make and interrupt the primary current at zero points of the first and last half cycle of the exposure. The target-film distance is varied, depending upon all the other factors involved; it generally has a value between 0.5 meter and 2 meters.

Secondary radiation from the heavier parts of the human body emanates in all directions, causing a general fogging effect over the whole area of the film and therefore reducing the contrast so important for accurate diagnosis. To minimize this, the Potter-Bucky diaphragm, consisting of a series of parallel lead strips perpendicular to the film and separated from each other by non-absorbing strips, is made to move over the surface of the film during exposure. The lead strips effectively absorb cross-radiation (secondary radiation) and therefore make better diagnostic results possible. Table 1 gives approximately the techniques required for making roentgenograms of various parts of the human body.

Table 1. Technique for Roentgenography and Roentgenoscopy

Purpose	Focal Spot-film Distance, meters	X-ray Tube Peak Voltage, kvp	X-ray Tube Current Average, ma	Exposure Time, seconds	Intensifying Screens
General roentgenoscopy..........	0.5–1.0	60–90	2–10	20–60	....
Roentgenography					
Hand........................	0.7–1.0	50–60	50	1.5	No
Elbow........................	0.7–1.0	40–60	50	1.5	No
Skull........................	0.7–1.0	60–75	100	1	Yes
Spine (use Bucky diaphragm)...	1.0	60–90	200	1	Yes
Colon (use Bucky diaphragm)...	1.0	60–90	100	1	Yes
Chest........................	1.25–2.0	45–85	30–500	0.4–0.033	Yes

17. APPARATUS

Roentgenoscopic apparatus comprises simply a high-tension transformer with the secondary connected directly to the x-ray tube (*self-rectification*) and a control apparatus for adjusting the x-ray-tube voltage and current to predetermined values from 60 to 90 kvp and from 2 to 10 ma. Portable roentgenographic machines are usually self-rectified. In order to avoid excessive voltage on the x-ray tube during the inverse half cycles, roentgenographic apparatus for use at 30 to 500 ma makes use of thermionic rectifiers. Up to 100 ma, a single thermionic rectifier is connected in series with the x-ray tube, resulting in half-wave rectification. At higher x-ray-tube currents (the usual ratings are 200 and 500 ma), four thermionic rectifiers are connected in a bridge circuit to produce full-wave rectifiers. In addition there are a few roentgenographic machines using three-phase rectifiers; there are also roentgenographic machines in which a high-voltage condenser with a capacitance of 0.25 to 1.0 μf is charged and subsequently discharged through the tube with effective exposure times of less than $^1/_{10}$ sec.

Several pieces of auxiliary equipment are essential for the production of good roentgenograms. Generally the film is contained in a light-tight cassette having a front plate of thin Bakelite or thin aluminum. Inside the cassette two intensifying screens, one on either side of the film, are arranged to maintain close contact with the film when the cassette is loaded. The x-rays passing through the screens excite fluorescence in the calcium tungstate or zinc sulfide contained therein, and this fluorescent light radiation records upon the film. Screens now used produce about 95 per cent of the total roentgenographic density, the other 5 per cent being caused by the direct absorption of x-rays in the film itself.

Conventional techniques require the use of 14-in. by 17-in. films for a roentgenogram of a chest. If it is desired to make chest films of many individuals, the cost is extremely high. This has led to the development of *photofluorographic* equipment for making x-ray surveys of the chest to discover early lesions of tuberculosis and other abnormalities. In this apparatus the x-rays which have traversed the patient impinge upon a fluorescent screen mounted in a light-tight box, opposite which a photographic camera is focused upon the screen. The camera records the image from the screen on 35-mm or 70-mm roll film, or on 4-in. by 5-in. flat film. Although the sharpness of the images of such photofluorograms is inferior to the sharpness of a roentgenogram, the results are believed by many radiologists to be accurate enough for surveys.

In order to standardize photofluorographic techniques, an automatic timing device has been developed. Light from a part of the screen is focused upon an electron multiplier tube, the output of which is integrated and used to operate a switch to discontinue the exposure at the end of the time required for producing the proper density on the film. This automatic timing device is now being tested for possible application to standard roentgenographic procedures.

The techniques of medical roentgenography have been increasingly used and further developed in the examination of industrial products. Thus the equipment operating at 2 Mev is capable of producing a satisfactory record on x-ray film of steel as thick as 12 in. The gamma rays from radium are also utilized to make films of metal parts. These are extremely valuable methods of inspecting the industrial products in order to detect faults within the product without destroying the product.

BIBLIOGRAPHY

McNeill, Clyde, *Roentgen Technique.* Charles C. Thomas, Springfield, Ill. (1939).
Robertson, J. K., *Radiology Physics.* Van Nostrand (1941).
St. John, A., and H. R. Isenburger, *Industrial Radiology.* 2nd Ed., Wiley (1943).
Weyl, C., S. R. Warren, Jr., and D. B. O'Neill, *Radiologic Physics.* Charles C. Thomas, Springfield, Ill. (1941).

18. MISCELLANEOUS DEVICES

Of the many electrical devices used in medicine that are not mentioned in the sections above, three pieces of apparatus are briefly described in this section.

The *electron microscope* is a device in which a beam of electrons traverses a thin sample of tissue or other material and impinges upon a fluorescent screen or film to produce an enlarged image representing a pattern of the specimen. The electron beam is controlled by means of electrostatic or electromagnetic lenses. The RCA type EMU is a recently developed commercially available electron microscope. The magnification can be varied from 100 to 20,000; the resolving power is somewhat less than 100 A. The device operates with a maximum difference of potential of 50 kvp. Films made on the electron microscope can be photographically enlarged so that the overall magnification may exceed 100,000. Among the limitations of the device are: (1) it is extremely difficult properly to prepare specimens; and (2) the specimen may be destroyed or seriously modified in structure by the electron beam.

The *betatron* is a device for accelerating electrons to velocities closely approaching the speed of light. It consists of a laminated iron core with an air gap. Between the poles of the gap there is mounted a doughnut-shaped vacuum tube into which pulses of electrons with a duration of about 2 μsec and velocities corresponding to approximately 80 kvp can be injected into the evacuated space. A winding on the core is fed from a source of 180-cycle alternating current in order to produce across the air gap a magnetic field in which the flux density varies sinusoidally with time; the first rising one-quarter cycle of the flux ($^1/_{720}$ sec) actuates the device. The pulse of electrons is injected into the doughnut a few microseconds after the magnetic flux has passed through zero and is increasing. Subsequently, the increasing magnetic field causes the electrons to be accelerated; they traverse the circular path within the tube several hundred thousand times in $^1/_{720}$ sec. It is possible so to shape the pole pieces that these electrons will remain in a stable circular orbit during their acceleration. At the end of the period ($^1/_{720}$ sec) they are caused to spiral outward from the stable orbit and to impinge upon a target, causing the production of x-rays. Betatrons for producing electron velocities corresponding to 20 Mev have been in use for several years, and an experimental model of a 100 Mev betatron has been announced by the General Electric Company. At high equivalent voltages, some of the phenomena observed in cosmic-ray studies have been produced in the laboratory for the first time.

A *cyclotron* is a device for accelerating protons and heavier positive ions. Its operation depends upon two facts: (1) charged particles traveling at right angles to a constant magnetic field traverse a circular path with constant linear velocity in the absence of an electric field; (2) charged particles are accelerated in the presence of an electric field in the direction of their motion. Near the center of a shallow, cylindrical evacuated cavity, positive ions are emitted from a suitable source at very low velocities. The cylindrical space is surrounded by two semicircular *dees*, and the source of ions is located in the gap between the dees. A source of high-frequency alternating voltage is applied to the dees so that those positive ions that are emitted during the peak of a particular half cycle are accelerated toward the dee which, at that instant, is negatively charged. Upon entering the dee the effect of the electrostatic field becomes negligibly small, and the electrons would continue in a straight line except that the dees are mounted in the gap between the poles of an electromagnet excited by direct current. Thus the ions that have been initially accelerated by their first traverse of the gap between the dees travel in a semicircle inside the dees until they again reach the gap. The high-frequency source of potential between the dees is adjusted so that the ion will be again accelerated as it crosses the gap the second time.

This process is repeated, and the radius of the semicircle increases each time the ions are accelerated by the electric field. Thus the ions spiral outward and are subjected to two increases in acceleration during each revolution. After several hundred revolutions the ions spiral outward to the maximum diameter of the dees and are then emitted tangentially through a window. By placing various materials in the path of the beam outside the cyclotron, important nuclear experiments can be carried out. For example, when protons are accelerated in the cyclotron and permitted to impinge on beryllium, neutrons are produced in relatively large quantities. The largest of these devices, 184 in. in diameter, was constructed at the University of California under the direction of Dr. E. O. Lawrence, the inventor of the apparatus. The cyclotron has been used for many important experiments in the phenomenal developments of nuclear physics. Note that the operation of the cyclotron in its simplest form depends upon the following fact: The time required for ions to move through one semicircle remains constant regardless of the linear velocity as long as the velocity is not comparable with the velocity of light. Ions in cyclotrons with equivalent voltages of 20 Mev or less satisfy the condition.

Several suggestions have been made of methods for accelerating both electrons and positive ions to velocities corresponding to voltages of 10^8 or 10^9 volts, and development of these methods is now proceeding.

BIBLIOGRAPHY

Electrical Engineering Staff MIT, *Applied Electronics*. Technology Press MIT and Wiley (1943).

Kerst, D. W., A 20-million Electron-volt Betatron or Induction Accelerator, *Rev. Scientific Instruments*, Vol. 13, 387–394 (September 1942).

Lawrence, E. O., and M. S. Livingston, Production of High-speed Light Ions without the Use of High Voltages, *Phys. Rev.* (Series 2), Vol. 40, 19–35 (April 1, 1932).

Schiff, L. I., Production of Particles beyond 200 Mev, *Rev. Scientific Instruments*, Vol. 17, 6–14 (January 1946).

HIGH-VOLTAGE SHOCK AND X-RAY BURN

19. HIGH-VOLTAGE SHOCK

The physiological effects of a high-voltage electric shock may be classified in two groups: the major effects, such as cessation of respiration or heart action; and the less serious effects, such as fractures and internal injuries due to falls, and also burns.

The first necessary action is to remove the victim from the circuit *without touching him with bare hands*. If it is not possible to open the circuit by means of a switch near at hand, then it is usually effective to move the conductor, or the victim, with a non-conductor. A physician should be summoned immediately, and it is important to administer first aid or artificial respiration immediately while awaiting his arrival. If the victim is breathing, heart stimulants may be administered hypodermically. The body should be rubbed to produce external warmth, and the clothing should be loosened in order to excite consciousness.

If the victim does not breathe, artificial respiration should be applied as follows: Lay the subject face down with arms and legs extended, and turn the face one side so that the mouth and nose are free for breathing. Remove foreign bodies such as tobacco, gum, and false teeth from the mouth, and have an assistant draw the subject's tongue forward. Kneel, straddling the subject's thighs, facing his head; rest the palms of the hands on the muscles of the small of the back with fingers spread over the lowest ribs. With arms held straight, swing forward slowly so that weight is gradually brought to bear upon the subject. This operation should take 2 or 3 sec. Immediately swing backward, removing the pressure. Repeat this procedure 12 or 15 times a minute, a complete respiration in 4 or 5 sec. Continue artificial respiration at least an hour without interruption or until the physician arrives. Do not give any liquid by mouth until the subject is fully conscious. After the victim breathes again, shock treatment may be administered as outlined above. If any bones have been fractured or if the victim appears to have received internal injury do not move him any more than is necessary and prepare for removal to the nearest hospital.

If the victim has received burns the surface of the skin should be protected from the air. Cut around any clothing that sticks to the burns and saturate adhering cloth or cotton dressing with $^1/_2$ per cent solution of picric acid or a solution of baking soda, about 1 teaspoonful to a pint of water.

BIBLIOGRAPHY

Dalziel, C. F., and J. B. Lagen, Effect of Electric Currents on Man, *Elect. Eng.*, Vol. 60, 63–66 (February 1941).

Ferris, L. P., B. G. King, P. W. Spence, and H. B. Williams, Effects of Electric Shock on the Heart, *Elect. Eng.*, Vol. 55, 498 (1936).

20. X-RAY BURN

As noted under the section on roentgen therapy the effect of x-rays on living tissue is always destructive. This effect is cumulative; that is, the successive application of small doses may cause destruction of living tissues. Workers in x-ray laboratories should be properly protected from exposure by the installation of lead or lead glass protective shields about the x-ray tube. If the worker must be in the field of the x-rays, he should wear a protective lead rubber apron, hood, and gloves. X-rays produced by low x-ray-tube voltages are absorbed almost completely by the skin. To eliminate these x-rays, which have normally no useful effect in treatment or roentgenography, a filter of $^1/_2$ mm or 1 mm of aluminum is placed between the x-ray tube and the patient.

The first effects of x-ray burn are reddening and itching of the skin and falling hair. Later, open sores may develop, which may subsequently cause the destruction of large areas.

Standards of x-ray protection have been set up and internationally accepted. *These standards should always be observed rigorously.* It is necessary to make periodic tests by having workers carry small pieces of light-protected film with a narrow lead strip covering part of the outer casing. These films are carried over a period of several working days and are then developed to discover whether any fogging has occurred on those portions of the film not protected by the lead strip. Such fogging indicates the need of more effective protective measures. Ionization chambers, similar in shape and size to fountain pens, may be charged and worn by a worker for several hours, and the exposure in fractions of a roentgen may be measured by means of a calibrated electrometer. Although there is some controversy concerning the magnitude of the safe daily tolerance dose, a commonly accepted value is 0.05 roentgen per day.

BIBLIOGRAPHY

X-ray Protection. Bureau of Standards, Handbook 20.

Protection against X-rays and Gamma Rays, *Radiology*, Vol. 46, 57–76 (January 1946).

American War Standard Safety Code for the Industrial Use of X-rays, Z54.1–1946, Am. Standards Assoc., New York.

INDEX

NOTE. The double numbers refer to both section and page numbers. For example, 1-38 indicates page 38 of Section 1.

ELECTRICAL ENGINEERS' HANDBOOK

ELECTRIC POWER

Fourth Edition

SECTION
1. MATHEMATICS, UNITS, AND SYMBOLS
2. PROPERTIES OF MATERIALS
3. ELECTRIC CIRCUITS AND ELECTRIC LINES
4. PRINCIPLES OF ELECTROCHEMISTRY
5. MEASUREMENTS AND MEASURING APPARATUS
6. RESISTORS, RHEOSTATS, CAPACITORS, REACTORS, ELECTROMAGNETS, AND PERMANENT MAGNETS
7. BATTERIES
8. DIRECT-CURRENT MACHINES AND ROTARY ENERGY CONVERTERS
9. ALTERNATING-CURRENT GENERATORS AND MOTORS
10. TRANSFORMERS
11. POWER RECTIFIERS AND INVERTERS
12. SWITCHGEAR AND CONTROL EQUIPMENT
13. POWER STATIONS AND SUBSTATIONS
14. POWER TRANSMISSION AND DISTRIBUTION
15. LIGHTING AND HEATING
16. INDUSTRIAL APPLICATIONS OF MOTORS AND SERVOMECHANISMS
17. TRANSPORTATION
18. ELECTROCHEMICAL AND ELECTROTHERMAL PROCESSES
19. RURAL ELECTRIFICATION DISTRIBUTION SYSTEMS

ELECTRIC COMMUNICATION AND ELECTRONICS

Fourth Edition

SECTION
1. MATHEMATICS, UNITS, AND SYMBOLS
2. PROPERTIES OF MATERIALS
3. RESISTORS, INDUCTORS, AND CAPACITORS
4. ELECTRON TUBES
5. ELECTRIC CIRCUITS, LINES, AND FIELDS
6. PASSIVE CIRCUIT ELEMENTS
7. VACUUM TUBE CIRCUIT ELEMENTS
8. FREQUENCY MODULATION
9. PULSE TECHNIQUES
10. TRANSMISSION CIRCUITS
11. ELECTRICAL MEASUREMENTS
12. ACOUSTICS
13. ELECTROMECHANICAL-ACOUSTIC DEVICES
14. OPTICS
15. ELECTRO-OPTICAL DEVICES
16. SOUND REPRODUCTION SYSTEMS
17. TELEPHONY
18. TELEGRAPHY
19. FACSIMILE TRANSMISSION AND RECEPTION
20. TELEVISION
21. ELECTRONIC CONTROL EQUIPMENT
22. AIDS TO NAVIGATION
23. MEDICAL APPLICATIONS OF ELECTRICITY